CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS07.html

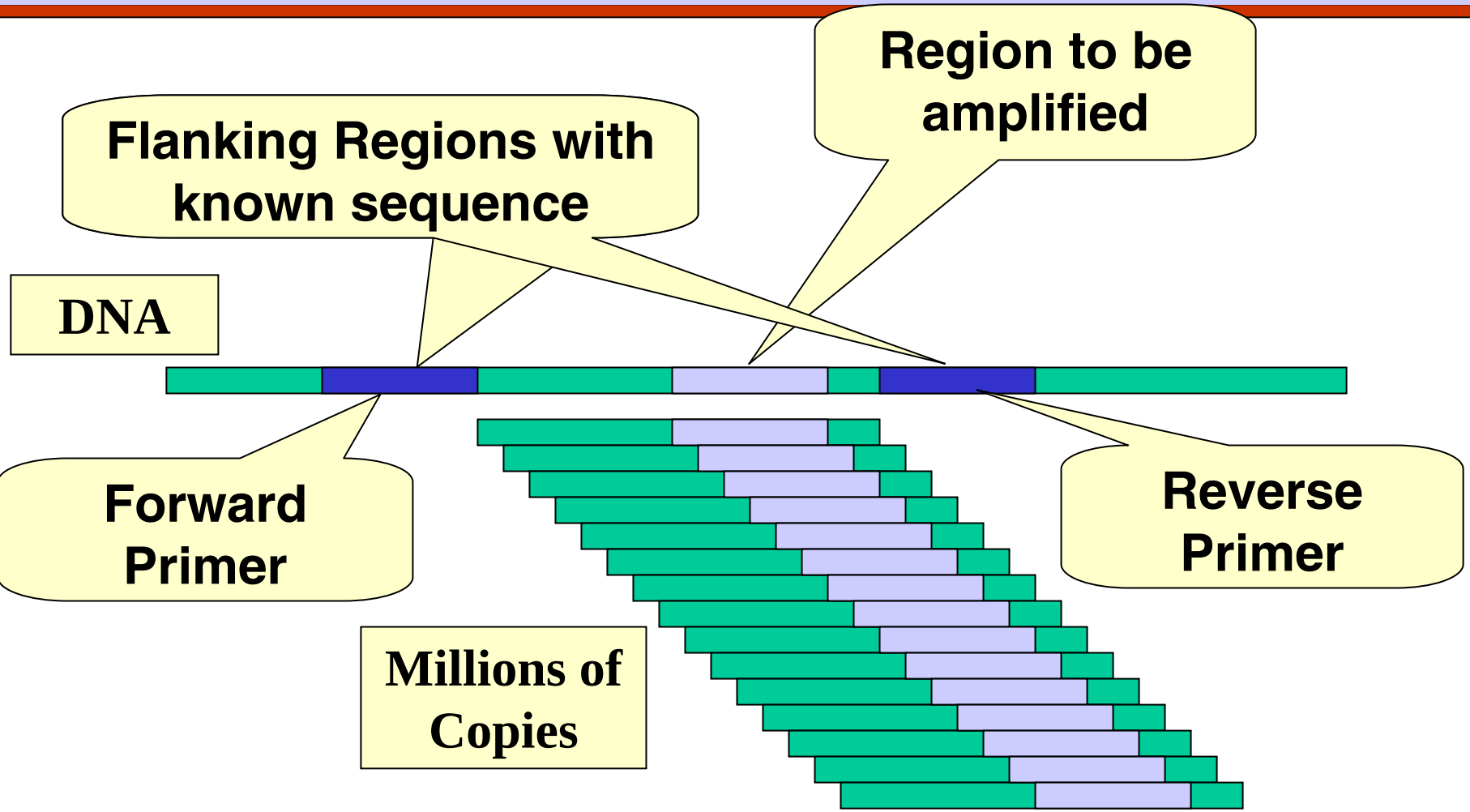
Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states

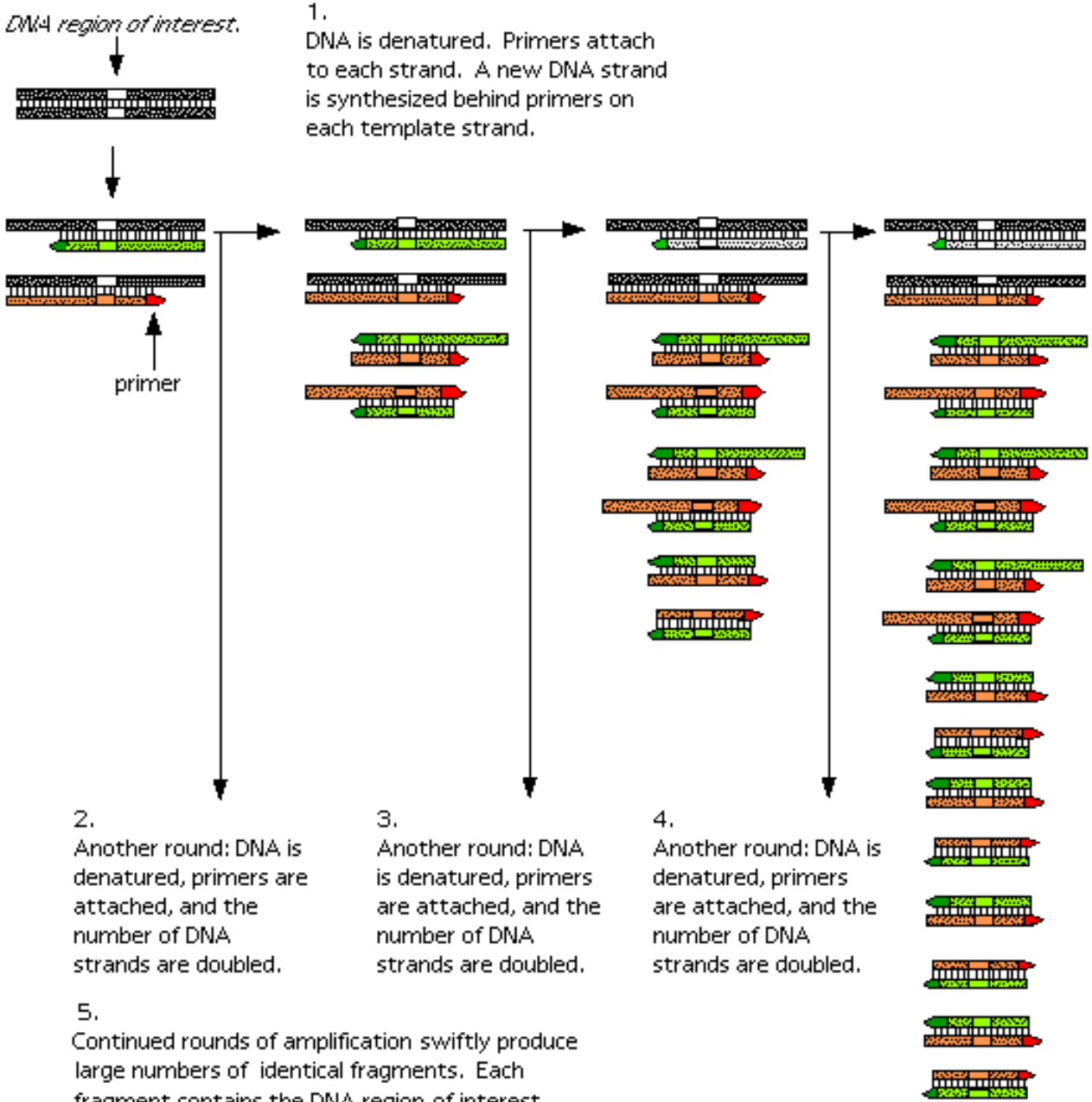
Polymerase Chain Reaction (PCR)

- For testing, large amount of DNA is needed
  - Identifying individuals for forensic purposes
    - (0.1 µL of saliva contains enough epithelial cells)
  - Identifying pathogens (viruses and/or bacteria)
- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences

PCR


Region to be amplified
Flanking Regions with known sequence
DNA
Forward Primer
Reverse Primer
Millions of Copies

POLYMERASE CHAIN REACTION



1.
DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.

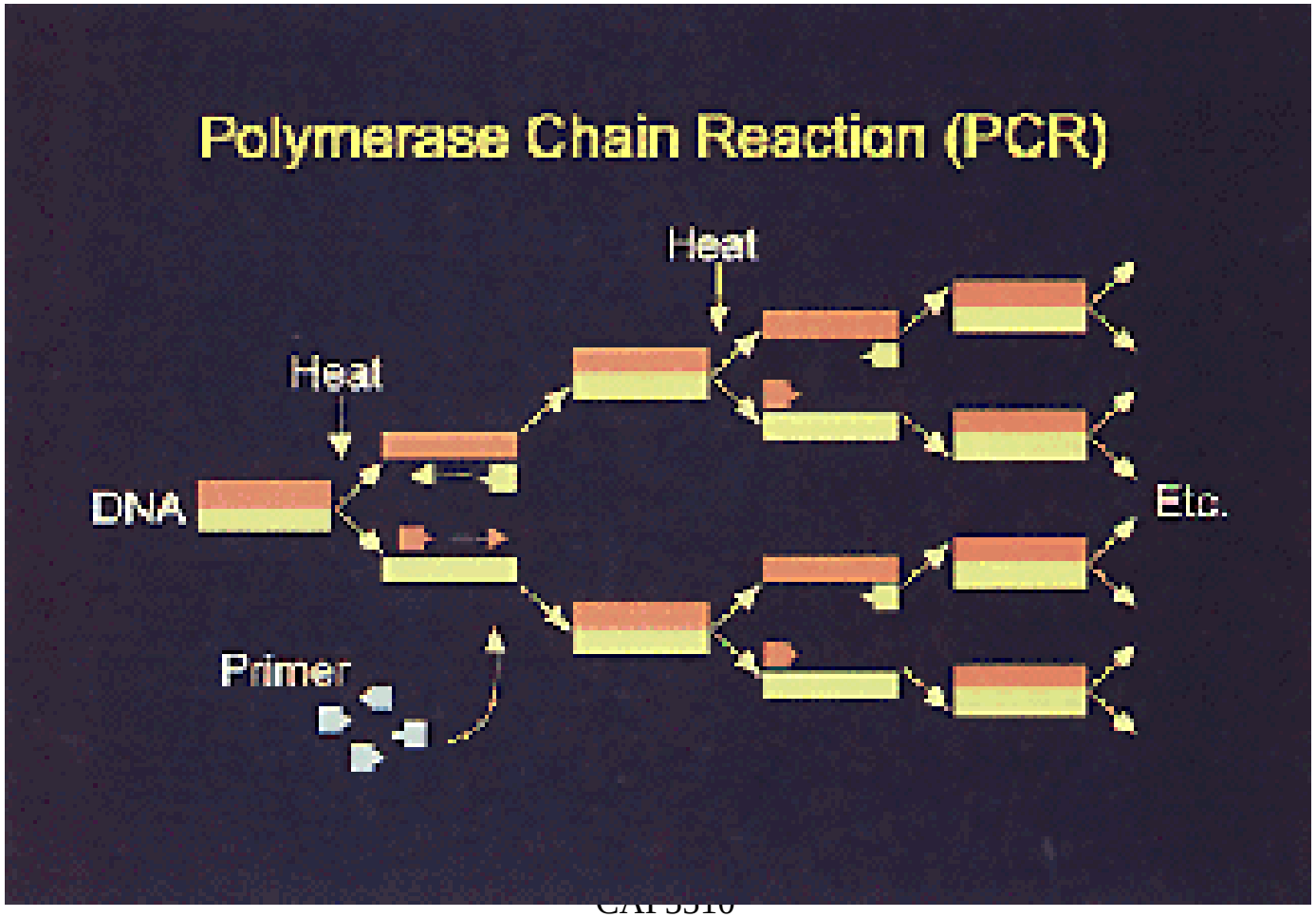
2.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

3.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

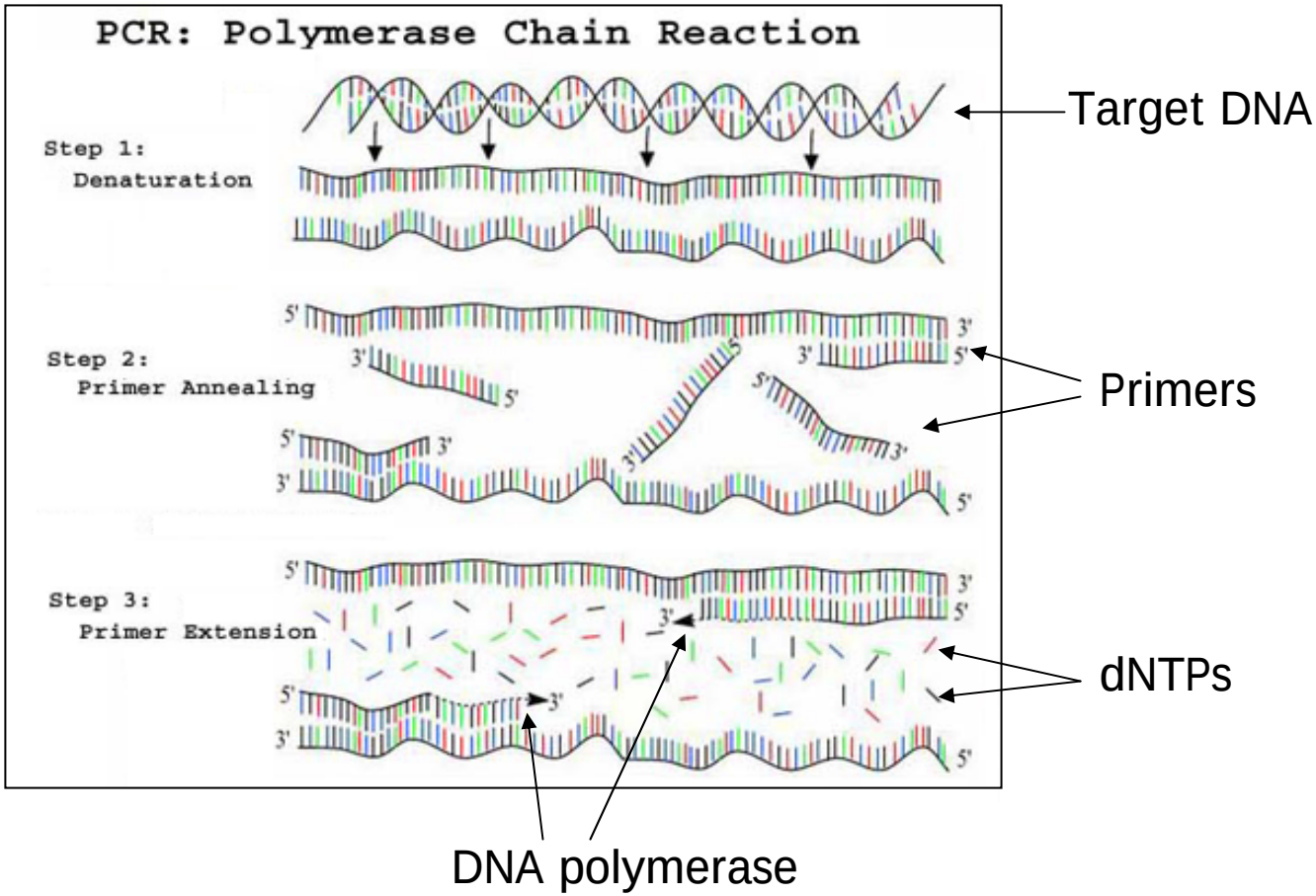
4.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

5.
Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.

PCR



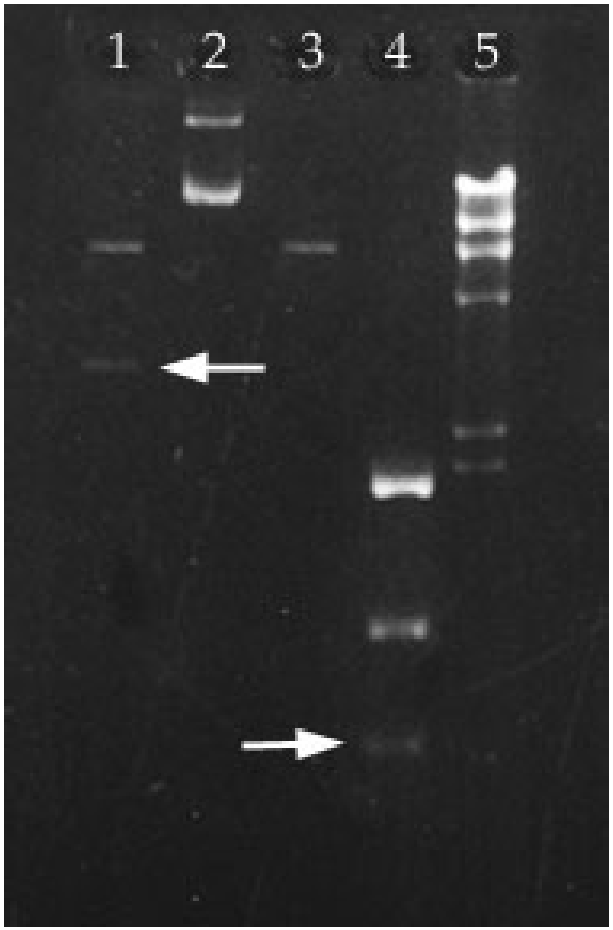
Schematic outline of a typical PCR cycle



Gel Electrophoresis

- Used to measure the lengths of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

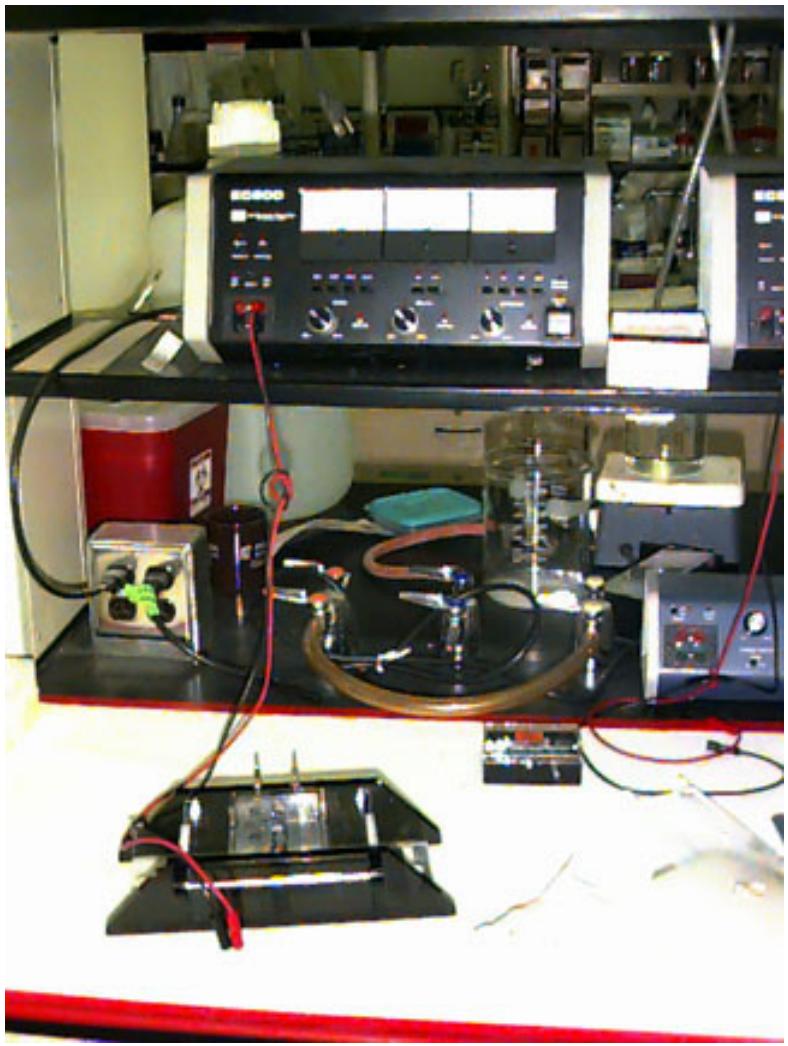
Gel Pictures

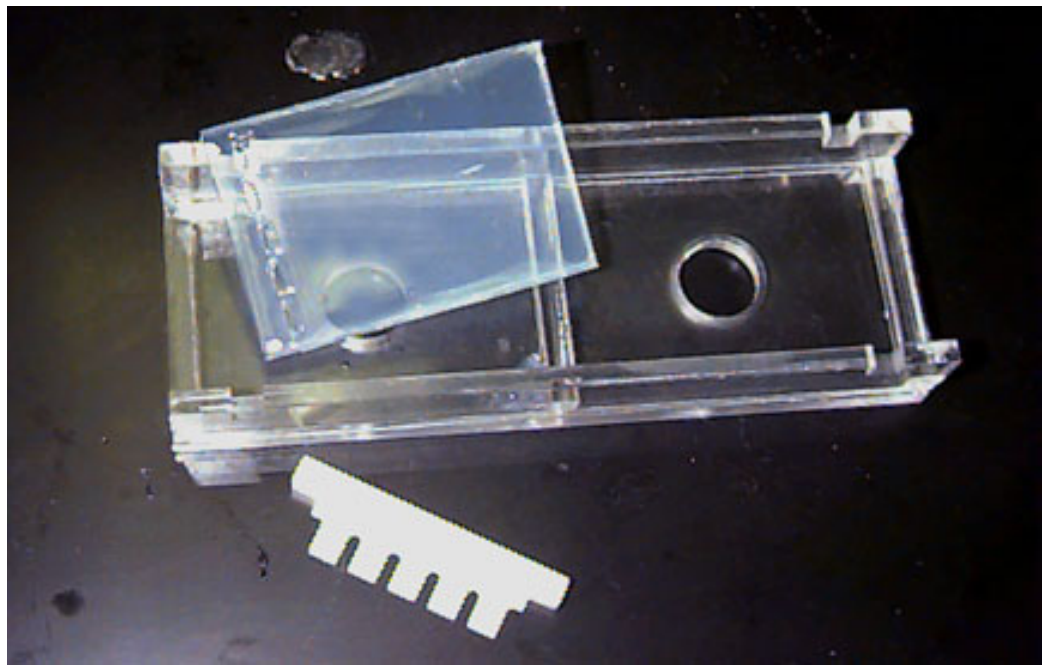


Gel Electrophoresis: Measure sizes of fragments

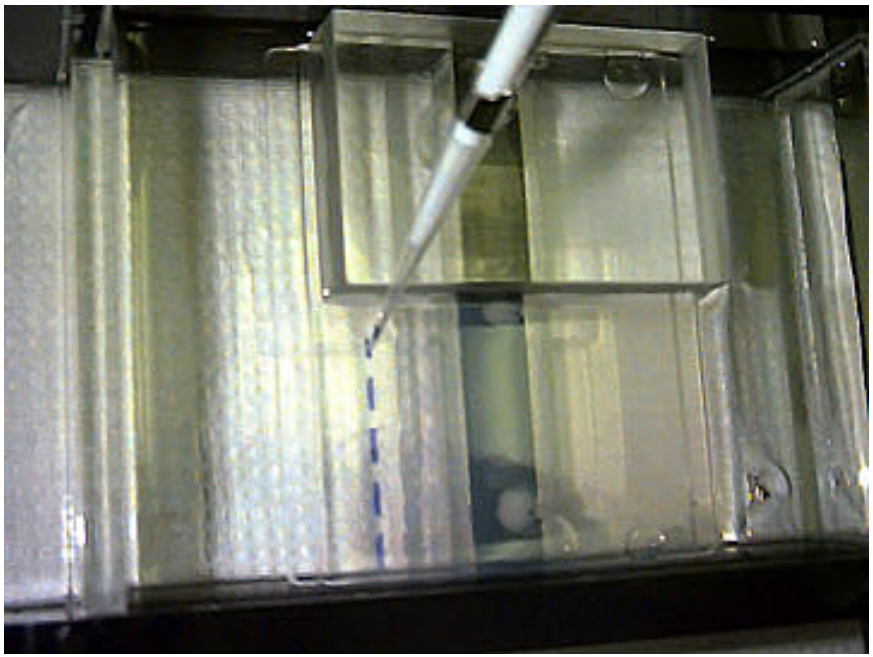
- The phosphate backbone makes DNA a highly negatively charged molecule. Thus DNA can be fractionated according to its size.
- Gel: allow hot 1 % solution of purifed agarose to cool and solidify/polymerize (like Jello).
- DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.
- Varying concentration of agarose makes different pore sizes & results.

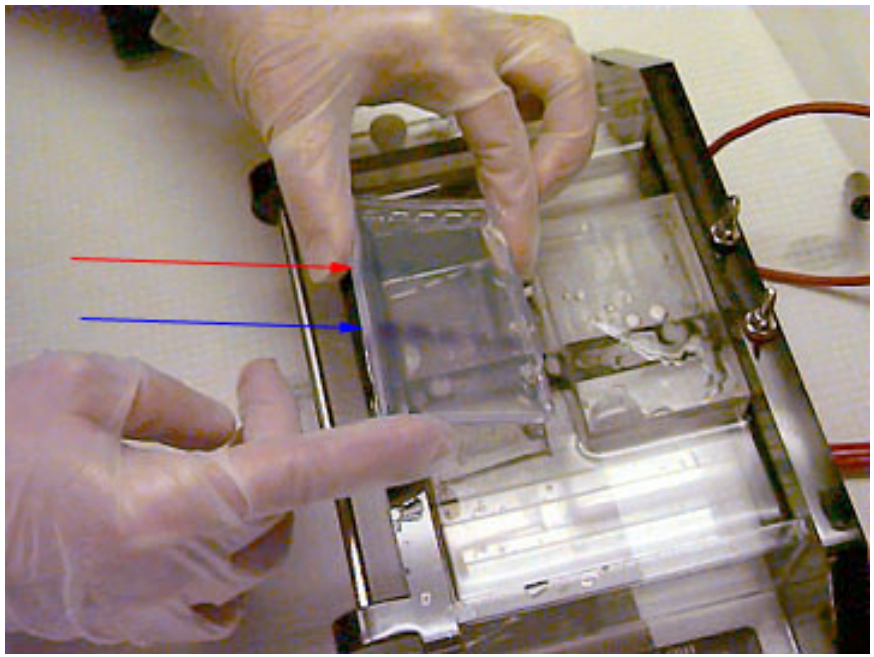
Gel Electrophoresis

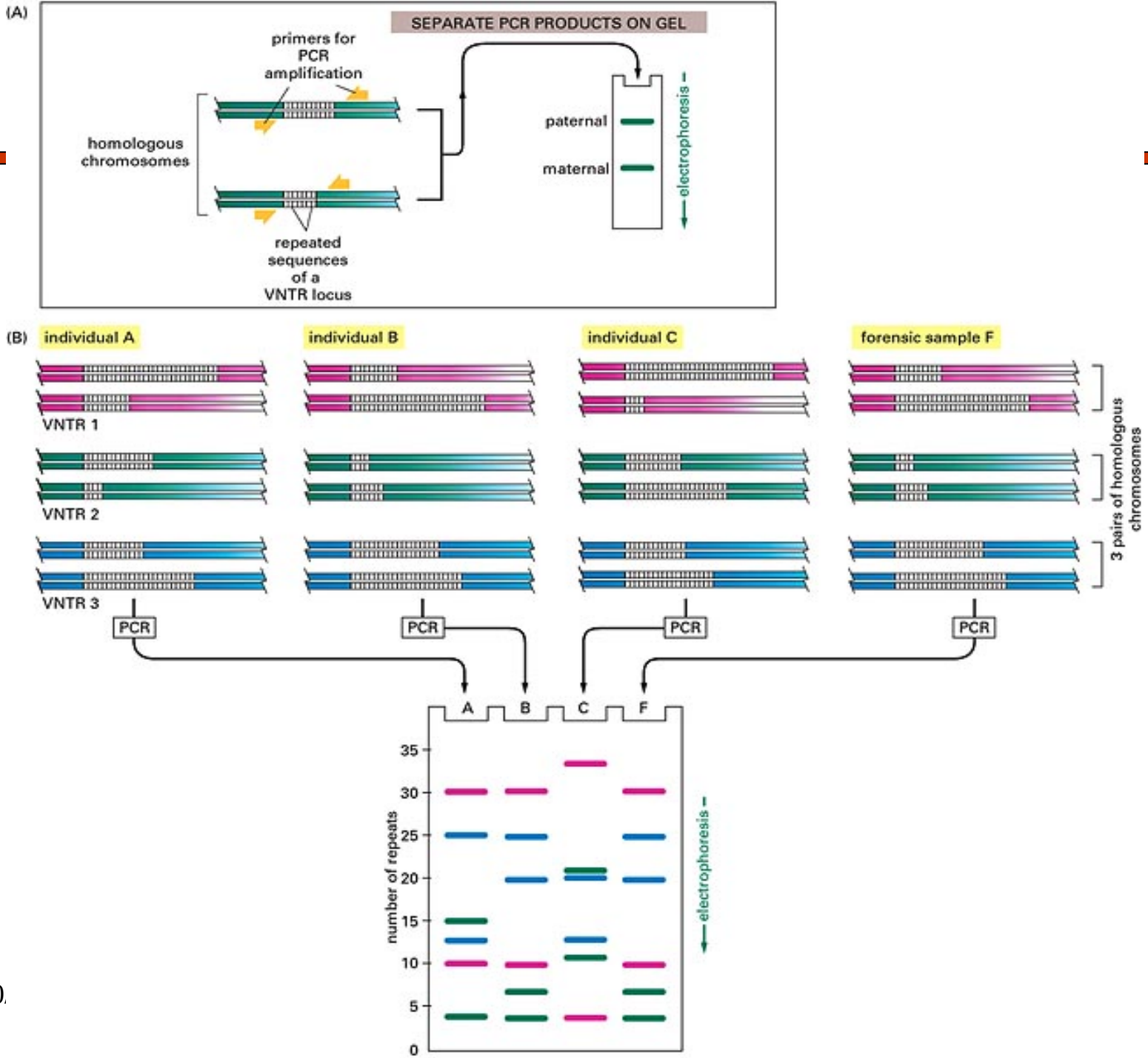




Gel Electrophoresis



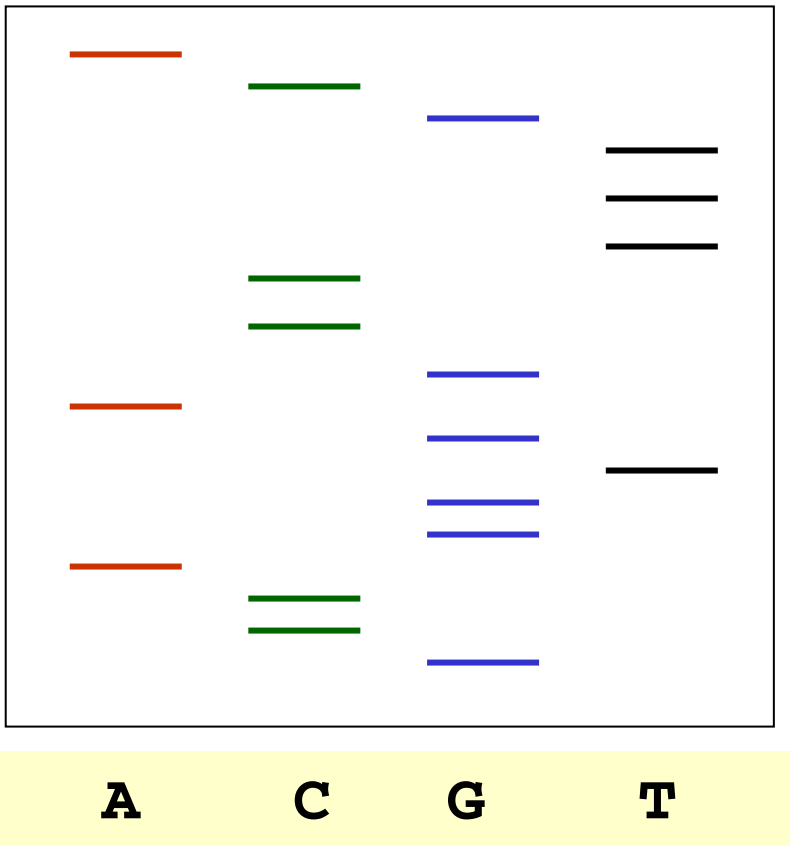



(A)
SEPARATE PCR PRODUCTS ON GEL
primers for PCR amplification
homologous chromosomes
repeated sequences of a VNTR locus
paternal
maternal
electrophoresis
(B)
individual A
individual B
individual C
forensic sample F
VNTR 1
VNTR 2
VNTR 3
3 pairs of homologous chromosomes
PCR
A
B
C
F
number of repeats
0
5
10
15
20
25
30
35
electrophoresis

Sequencing a Fragment Using Gels

- Isolate the desired DNA fragment.
- Using the "starving method" obtain all fragments that end in A, C, G, T
- Run gel with 4 lanes and read the sequence

Application of Gels: Sequencing



GCCAGGTGAGCCTTTGCA

Sequencing

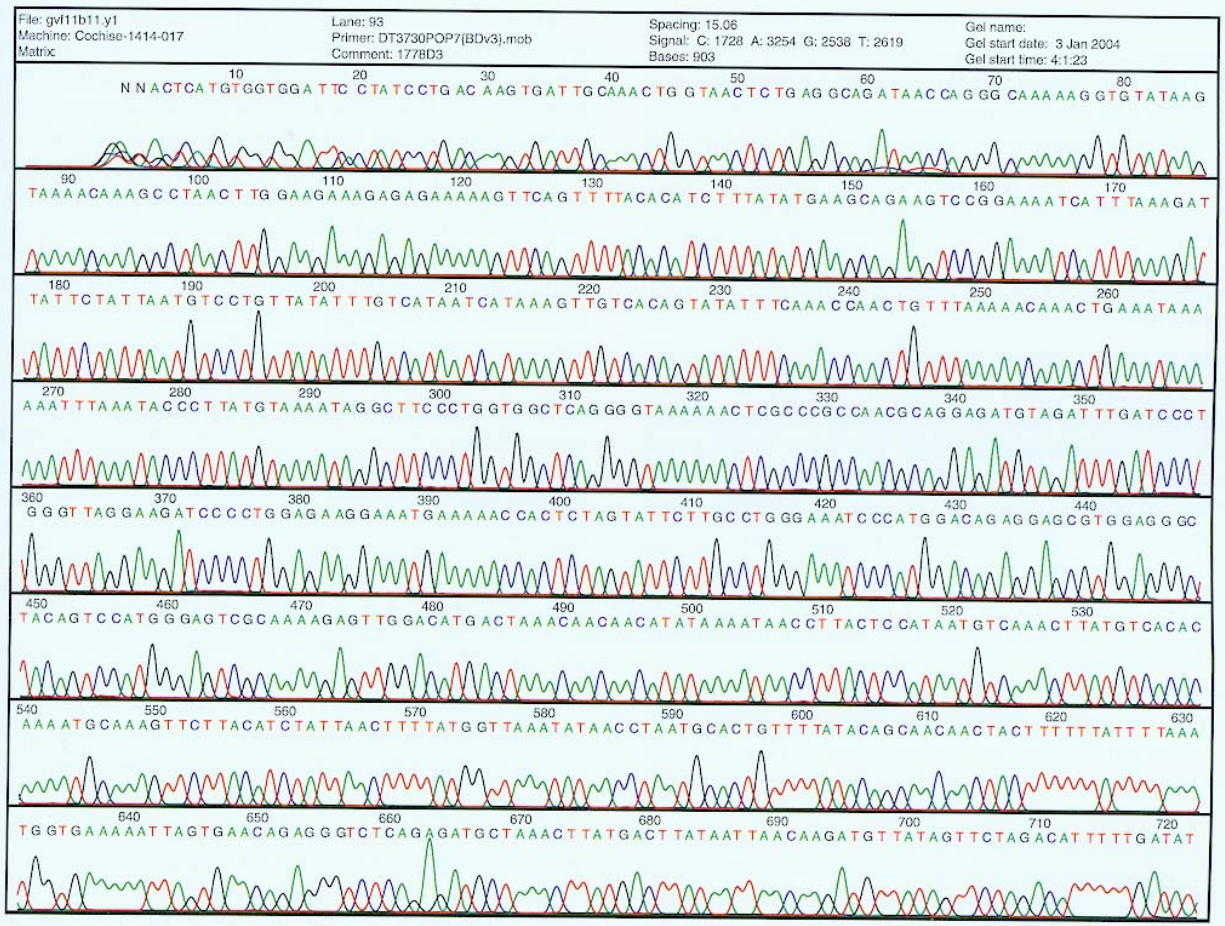
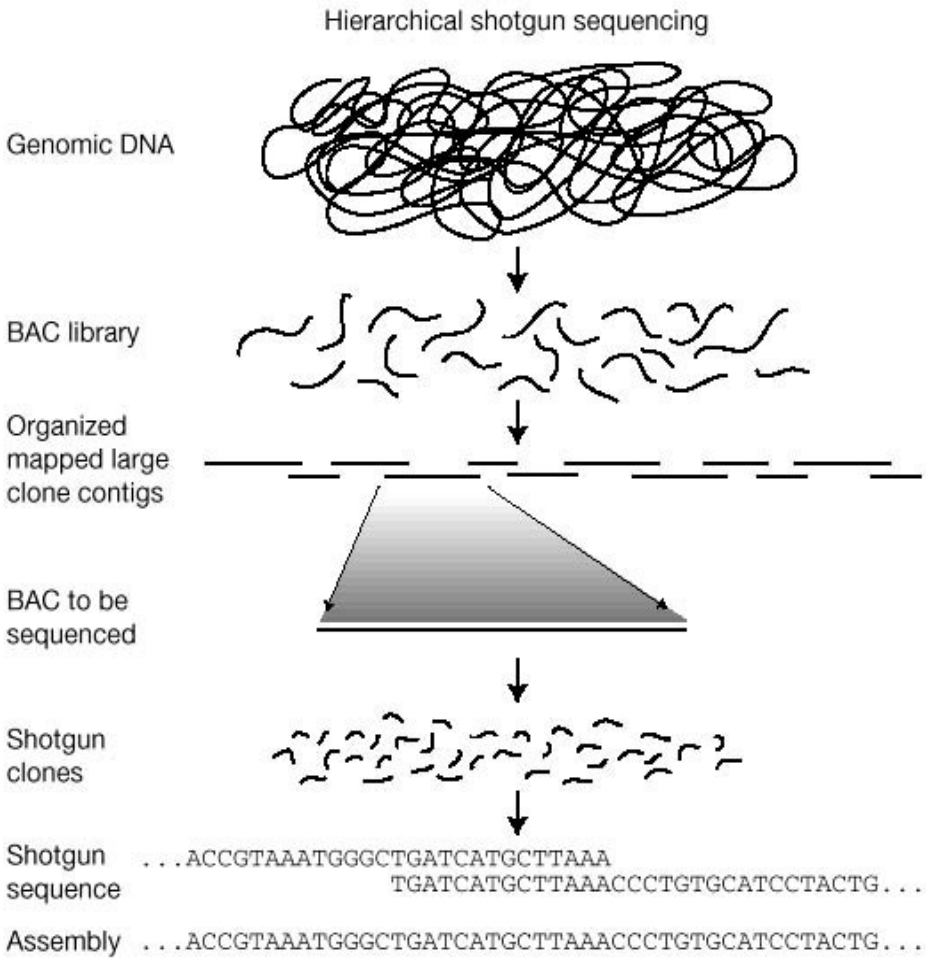


FIGURE 13.3 A sample chromatogram, as viewed with the vtrace program (Ewing, 2002). Signal intensities corresponding to fragments ending with A (green), C (blue), G (black), and T (red) are shown out to approximately 722 bases.

Sequencing

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Sequencing

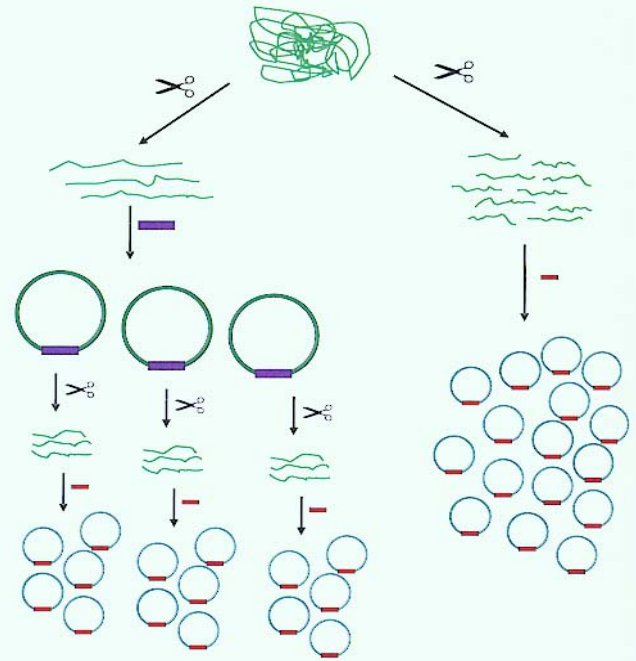
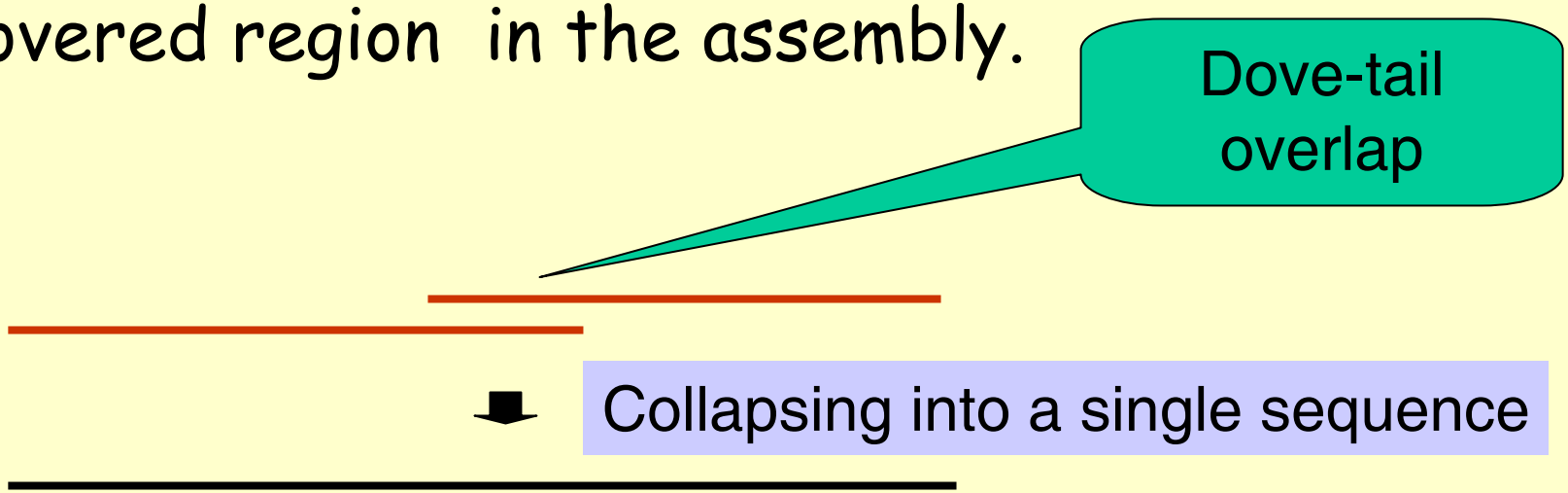


FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.

Sequencing: Generate Contigs

- Short for "contiguous sequence". A continuously covered region in the assembly.



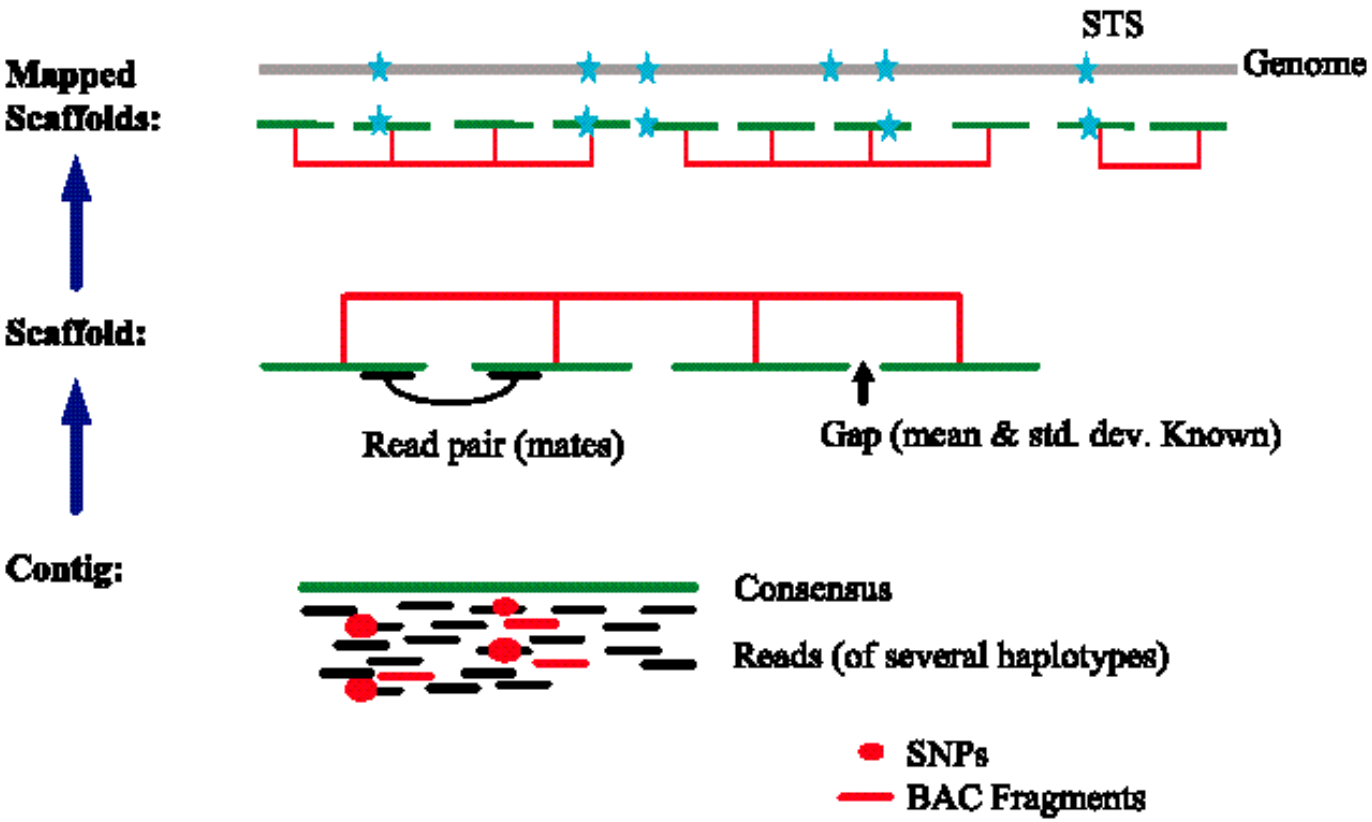
- Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
  Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

Supercontigs/Scaffolds

- A supercontig is formed when an association can be made between two contigs that have no sequence overlap.
  - This commonly occurs using information obtained from paired plasmid ends. For example, if both ends of a BAC clone are sequenced, then it can be inferred that these two sequences are approximately 150-200 Kb apart (based on the average size of a BAC). If the sequence from one end is found in a particular sequence contig, and the sequence from the other end is found in a different sequence contig, the two sequence contigs are said to be linked. In general, it is useful to have end sequences from more than one clone to provide evidence for linkage.

    [NCBI Genome Glossary]

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
- Then run it on your PC.

Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-

  - **--ACCGT--**
  - **----CGTGC**
  - **TTAC-----**
  - **-TACCGT-**
  - **TTACCGTGC**

Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage

Assembly: Complications

w = AGTATTGGCAATC

z = AATCGATG

u = ATGCAAACCT

x = CCTTTTGG

y = TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
--------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

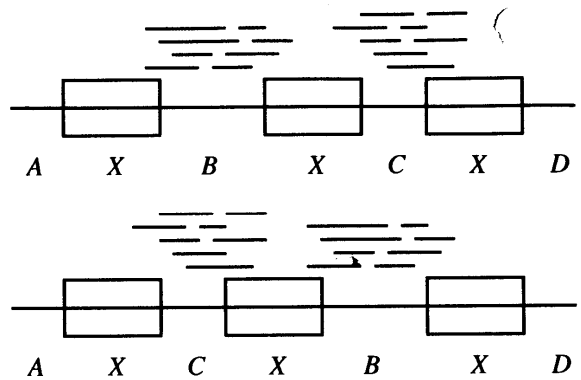


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

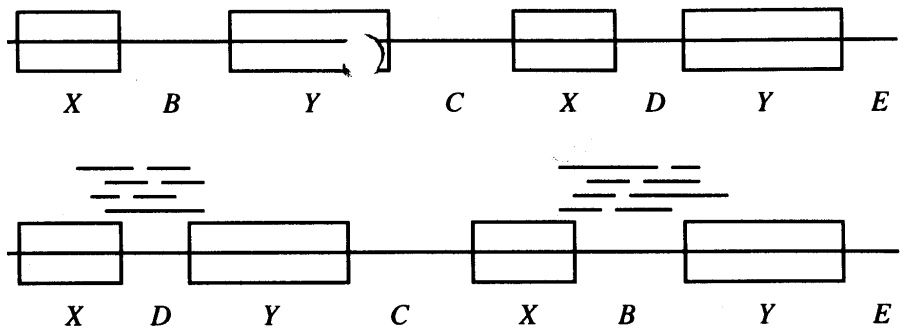


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

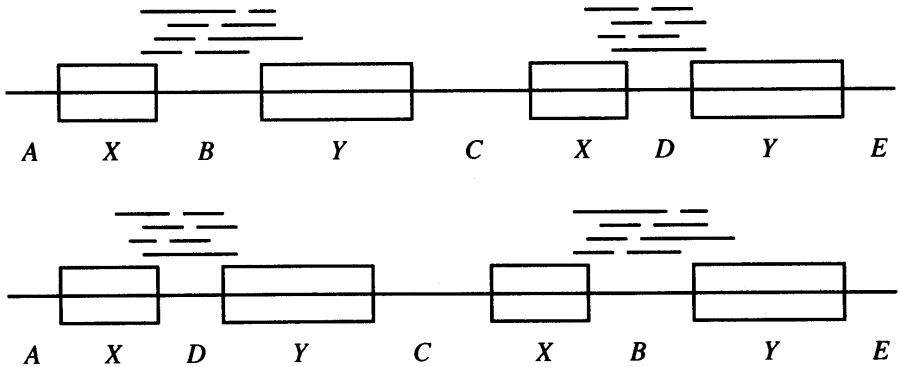


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

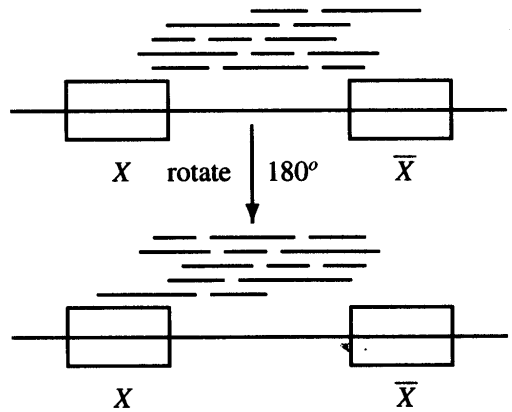


FIGURE 4.10

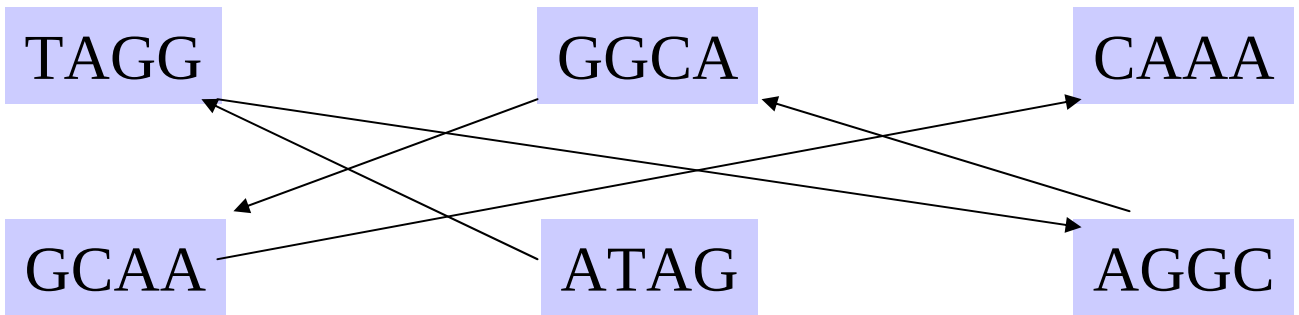
Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

Other sequencing methods

- Sequencing by Hybridization (SBH)
- Dual end sequencing
- Chromosome Walking (see page 5-6 of Pevzner's text).

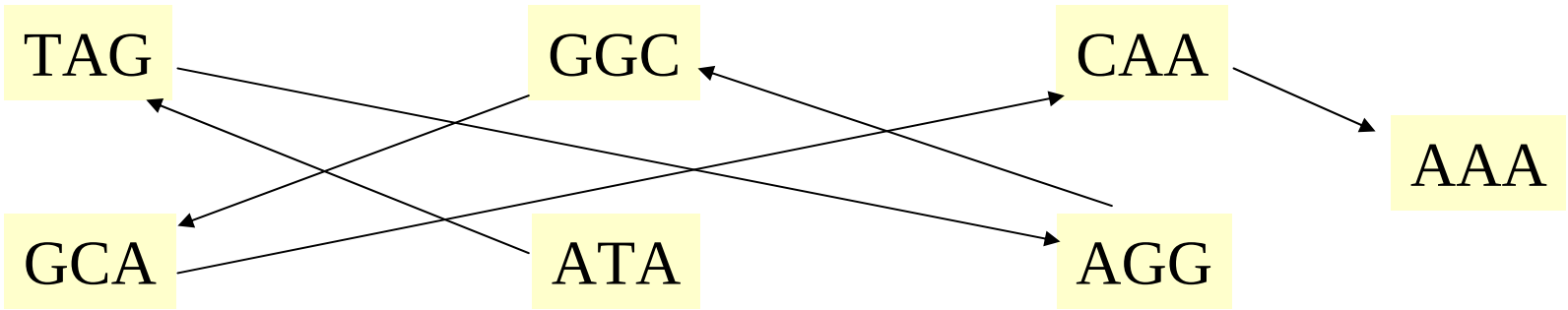
SBH

- Suppose that the only length 4 fragments that hybridize to S are: **TAGG, GGCA, CAAA, GCAA, ATAG, AGGC**. Then what is S, if it is of length ~9?



Hamiltonian Path Problem

SBH

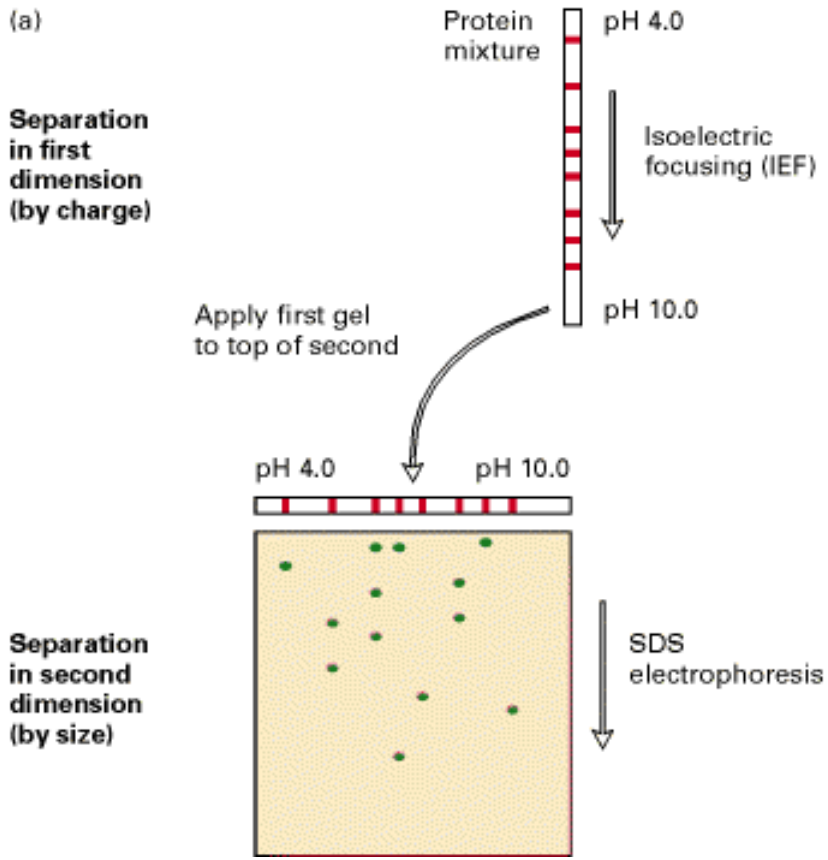


Eulerian Path Problem

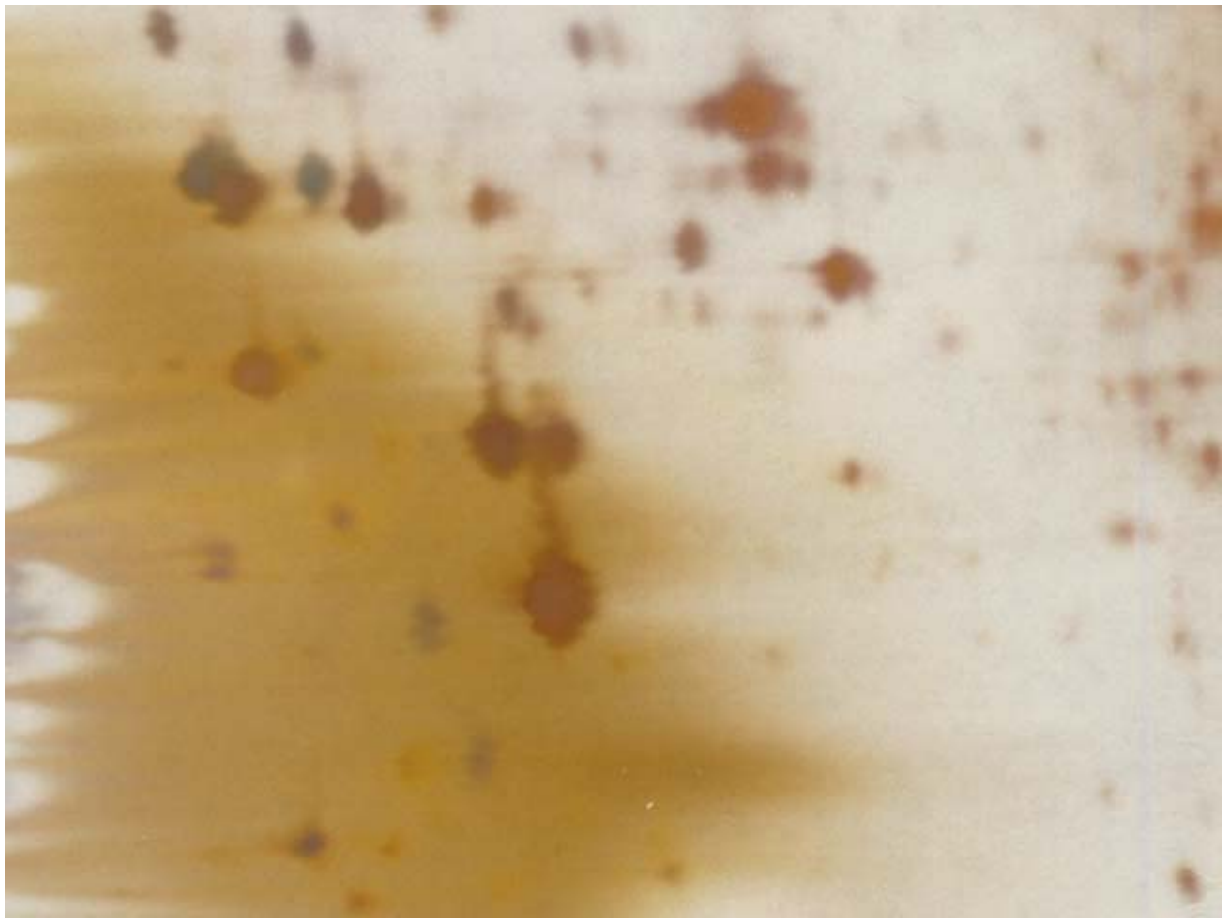
Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]
- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).
- Phrap (http://bozeman.genome.washington.edu/phrap.docs/phrap.html)
- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)
- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]
- GCG (http://www.gcg.com)
- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)
- SPACE, SPASS [Lawrence Berkeley Labs] (http://www-hgc.lbl.gov/inf/space.html)
- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)

2D-Gels


(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis



2D-Gels

First Dimension Methodology of a 2D Gel:

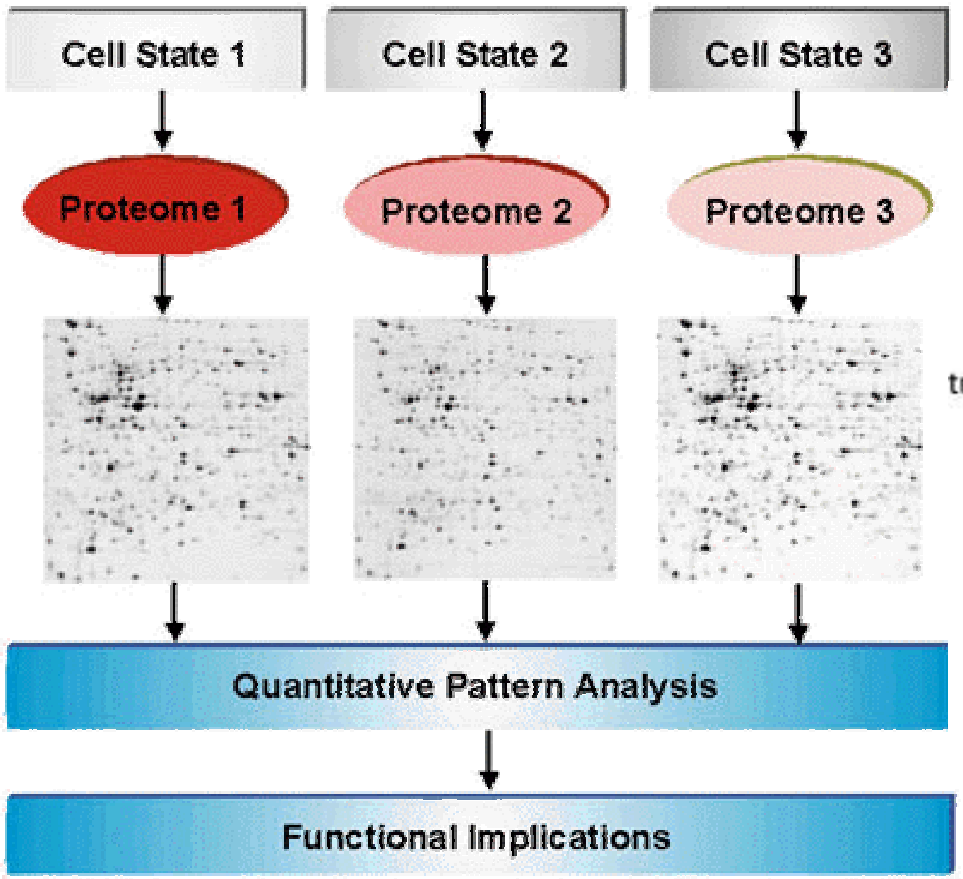
Denatured cell extract layered on a glass tube filled with polyacrylamide saturated with solution of ampholytes, a mixture of polyanionic[(-) charged] and polycationic [(+) charged] molecules. When placed in an electric field, the ampholytes separate and form continuous gradient based on net charge. Highly polyanionic ampholytes will collect at one end of tube, highly polycationic ampholytes will collect at other end. Gradient of ampholytes establishes pH gradient. Charged proteins migrate through gradient until they reach their pI, or isoelectric point, the pH at which the net charge of the protein is zero. This resolves proteins that differ by only one charge.

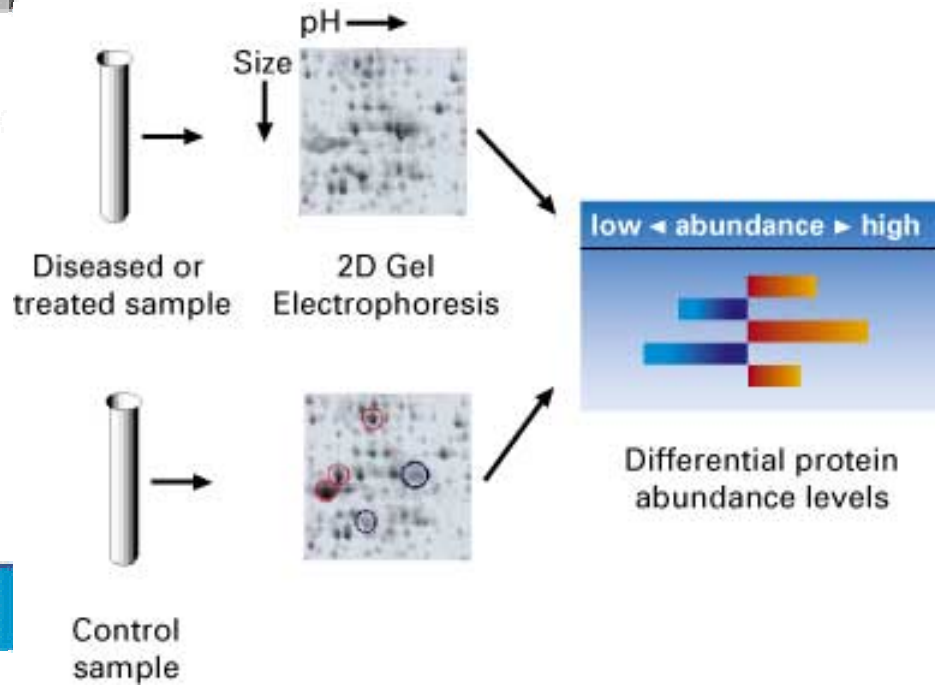
Entering the Second Dimension:

Proteins that were separated on IEF gel are next separated in the second dimension based on their molecular weights. The IEF gel is extruded from tube and placed lengthwise in alignment with second polyacrylamide gel slab saturated with SDS. When an electric field is imposed, the proteins migrate from IEF gel into SDS slab gel and then separate according to mass. Sequential resolution of proteins by their charge and mass can give excellent separation of cellular proteins. As many as 1000 proteins can be resolved simultaneously.

*Some information was taken from Lodish *et al.* Molecular Cell Biology.

2D-gels

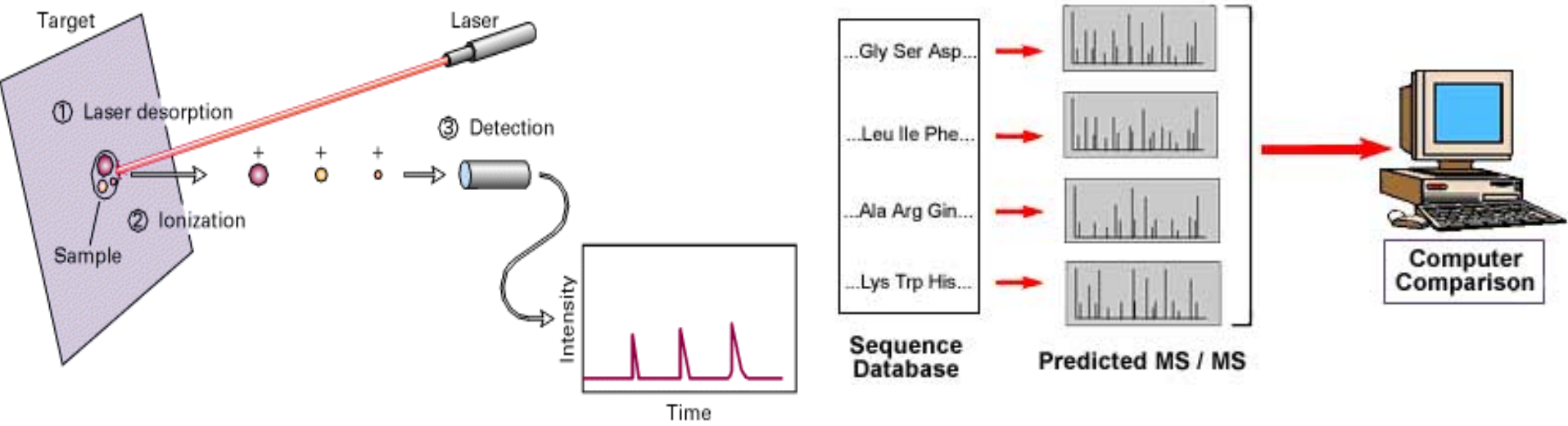




Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels

Mass Spectrometry



Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.
- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.

Genomics

- Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker: uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)

Genomics

- Study of all genes in a genome
  - Gene Expression
    - Microarray experiments & analysis
      - Probe design (CODEHOP)
      - Array image analysis (CrazyQuant)
      - Identifying genes with significant changes (SAM)
      - Clustering

Comparative Genomics

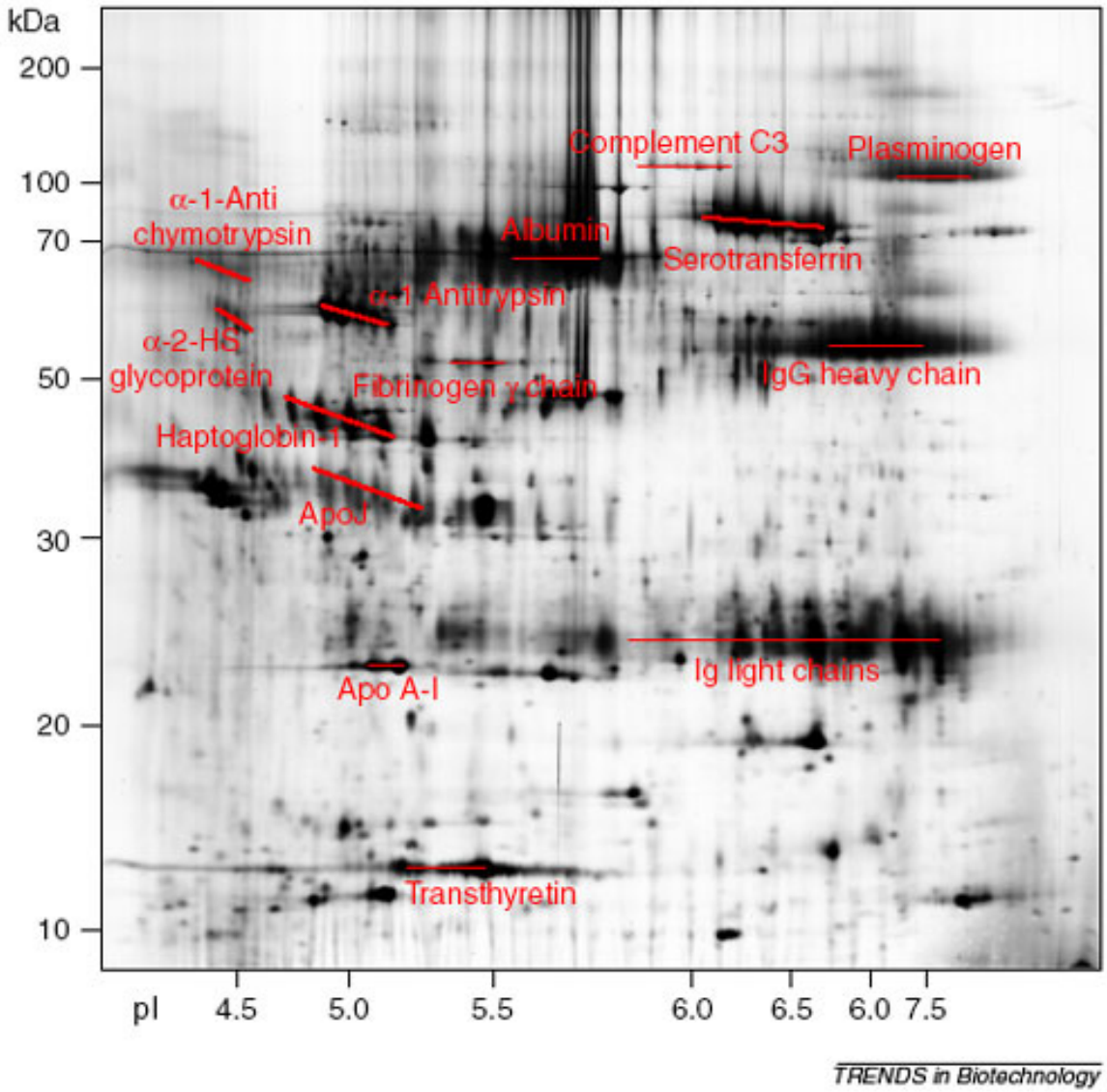
- Comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker, MultiPipMaker, EnteriX: PipMaker uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)
    - Many more!

Databases for Comparative Genomics

- PEDANT useful resource for standard questions in comparative genomics. For e.g., *how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.*
- COGs Clusters of orthologous groups of proteins.
- MBGD Microbial genome database searches for homologs in all microbial genomes

Proteomics

- Study of all proteins in a genome, or comparison of whole genomes.
  - Whole genome annotation & Functional proteomics
  - Whole genome comparison
  - Protein Expression: 2D Gel Electrophoresis


kDa
200
100
70
50
30
20
10
pI
4.5
5.0
5.5
6.0
6.5
6.0
7.5
Complement C3
Plasminogen
α-1-Anti
chymotrypsin
Albumin
Serotransferrin
α-1 Antitrypsin
α-2-HS
glycoprotein
Fibrinogen γ chain
IgG heavy chain
Haptoglobin-1
ApoJ
Ig light chains
Apo A-I
Transthyretin
TRENDS in Biotechnology

Other Proteomics Tools

From ExPASy/SWISS-PROT:

- **AACompIdent** identify proteins from aa composition

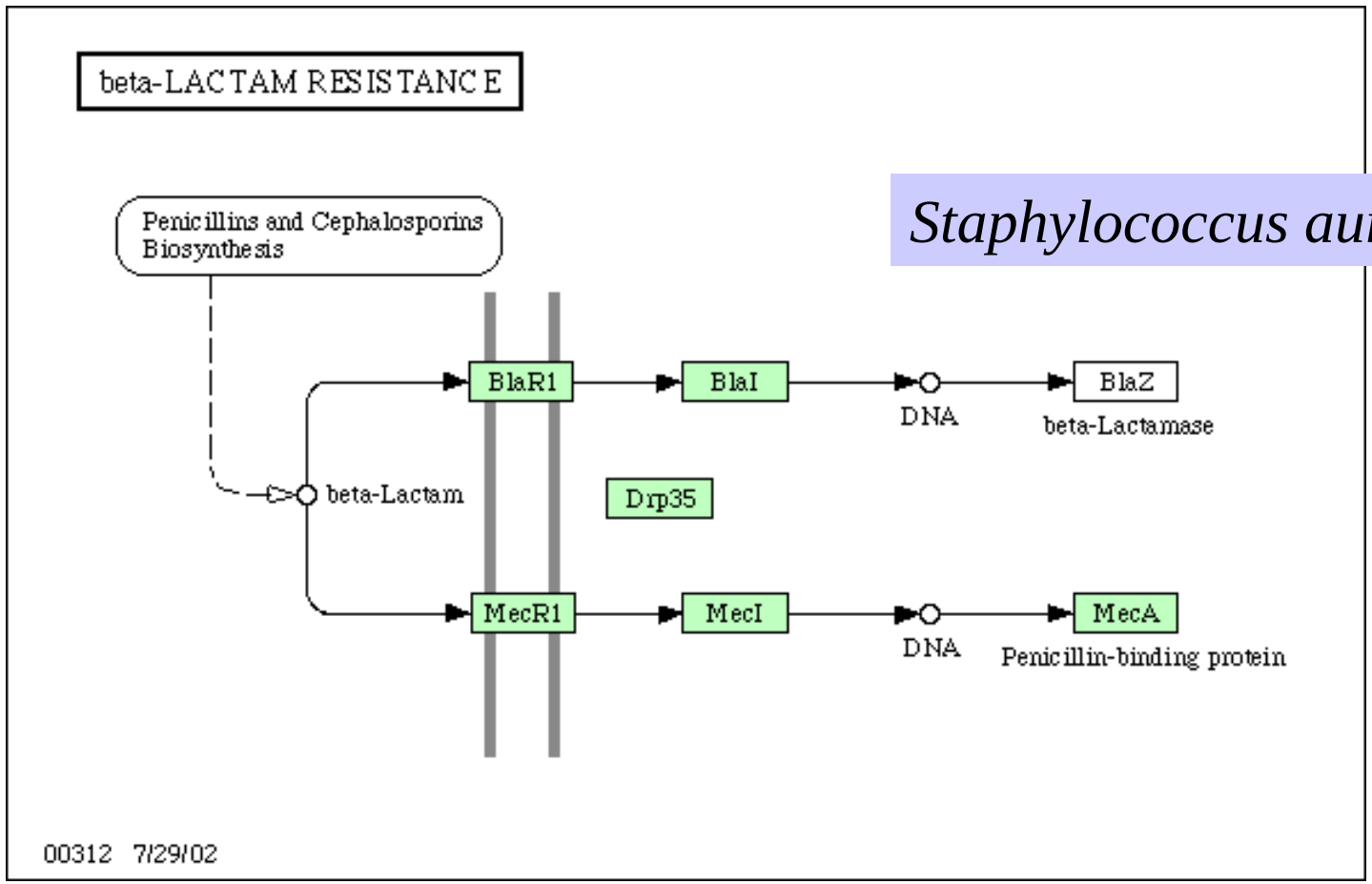
[Input: aa composition, isoelectric point, mol wt., etc. Output: proteins from DB]

- **AACompSim** compares proteins aa composition with other proteins
- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins
- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins
- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.
- **PeptideMass** theoretical mass fingerprint for a given protein.
- **GlycoMod** predicts oligosaccharide modifications from mass difference
- **TGREASE** calculates hydrophobicity of protein along its length

Gene Networks & Pathways

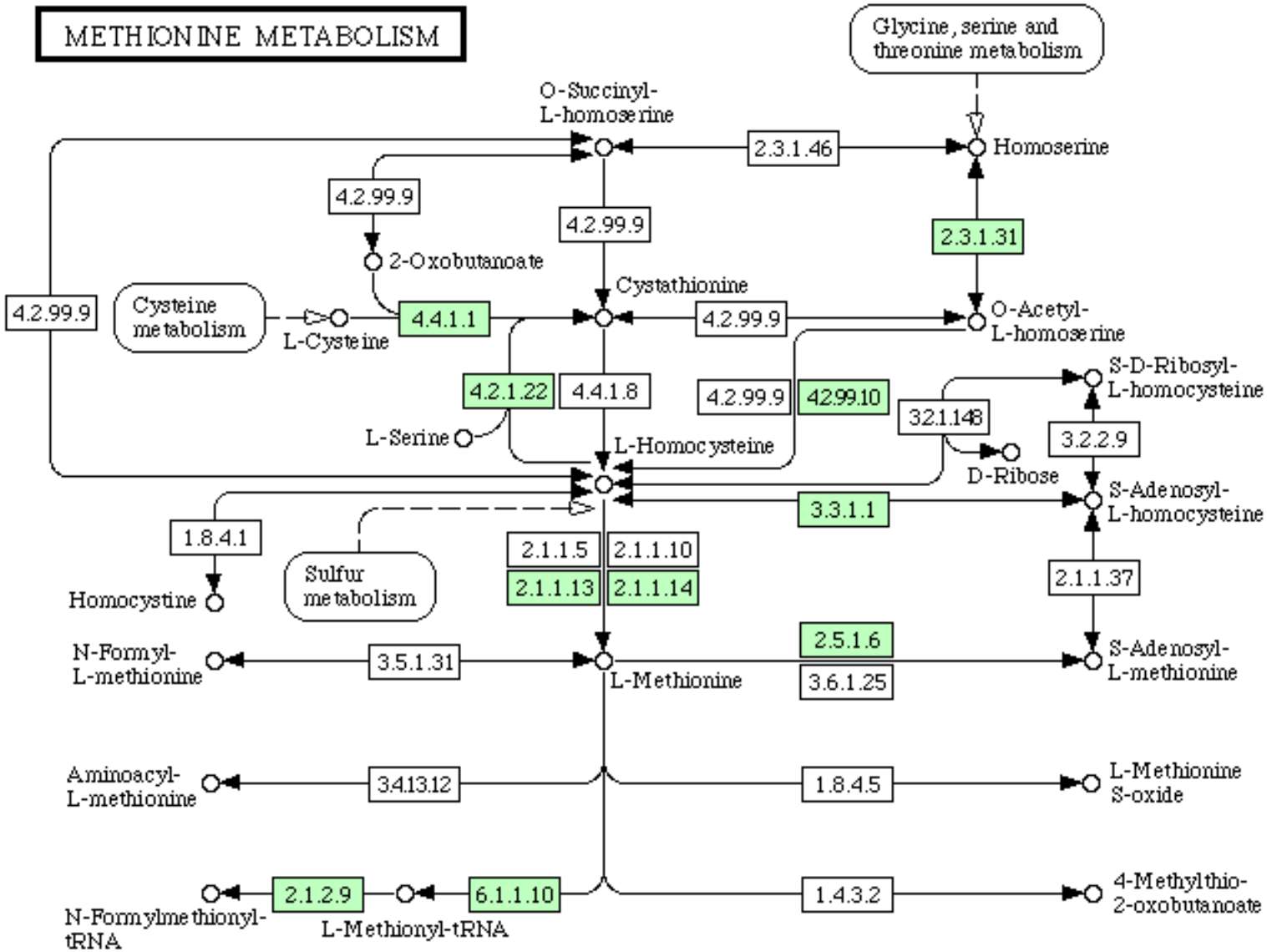
- Genes & Proteins act in concert and therefore form a complex network of dependencies.

Pathway Example from KEGG



Staphylococcus aureus

Pseudomonas aeruginosa



00271 8/6/02

STSs and ESTs

- Sequence-Tagged Site: short, unique sequence
- Expressed Sequence Tag: short, unique sequence from a coding region
  - 1991: 609 ESTs [Adams et al.]
  - June 2000: 4.6 million in dbEST
  - Genome sequencing center at St. Louis produce 20,000 ESTs per week.

What Are ESTs and How Are They Made?

- Small pieces of DNA sequence (usually 200 - 500 nucleotides) of low quality.
- Extract mRNA from cells, tissues, or organs and sequence either end. Reverse transcribe to get cDNA (5' EST and 3'EST) and deposit in EST library.
- Used as "**tags**" or markers for that gene.
- Can be used to identify similar genes from other organisms (Complications: variations among organisms, variations in genome size, presence or absence of **introns**).
- 5' ESTs tend to be more useful (cross-species conservation), 3' EST often in UTR.

DNA Markers

- Uniquely identifiable DNA segments.
- Short, <500 nucleotides.
- Layout of these markers give a map of genome.
- Markers may be polymorphic (variations among individuals). Polymorphism gives rise to alleles.
- Found by PCR assays.

Polymorphisms

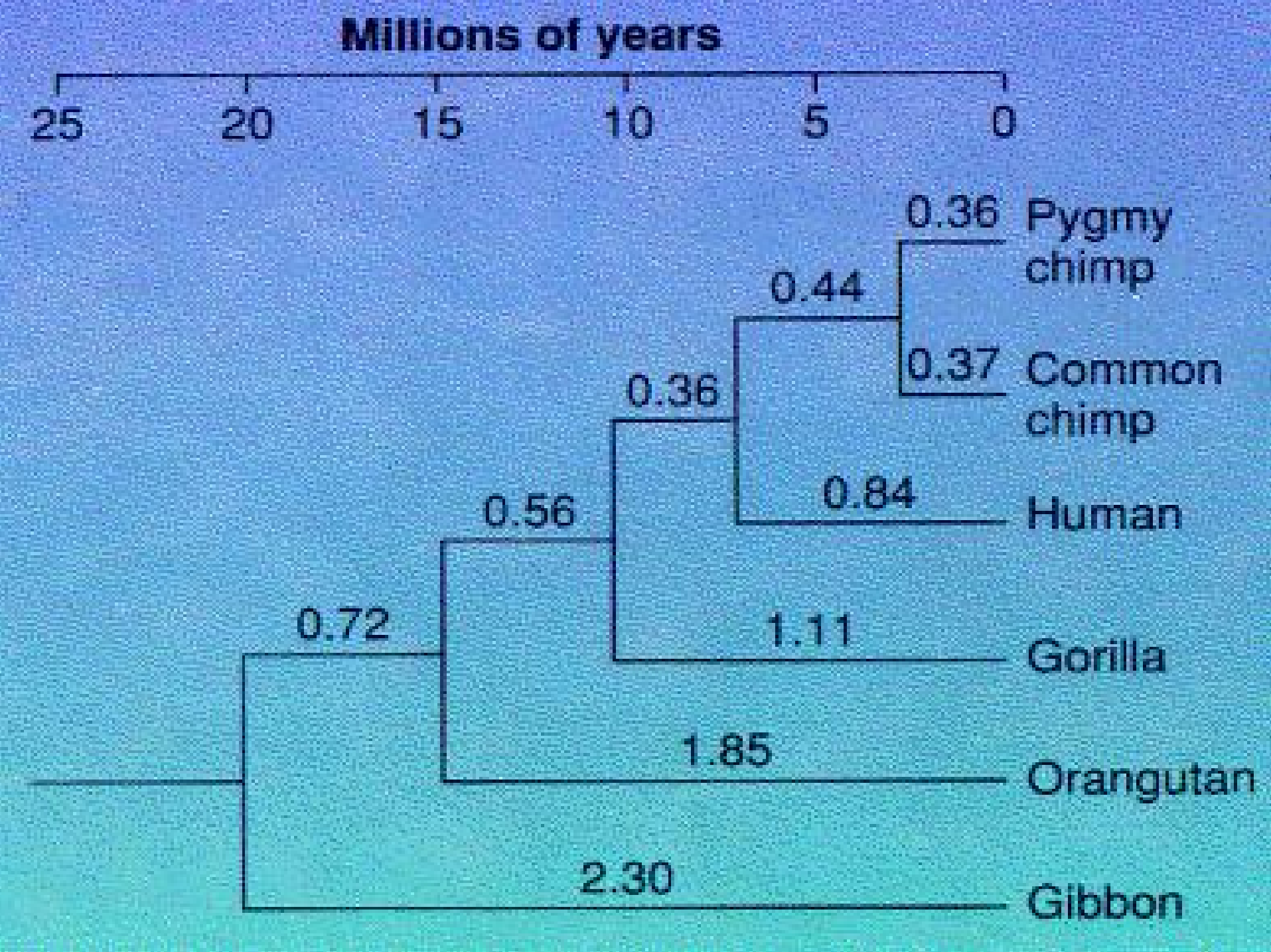
- Length polymorphisms
  - Variable # of tandem repeats (VNTR)
  - Microsatellites or short tandem repeats
  - Restriction fragment length polymorphism (RFLP) caused by changes in restriction sites.
- Single nucleotide polymorphism (SNP)
  - Average once every ~100 bases in humans
  - Usually biallelic
  - dbSNP database of SNPs (over 100,000 SNPs)
  - ESTs are a good source of SNPs

SNPs

- SNPs often act as "disease markers", and provide "genetic predisposition".
- SNPs may explain differences in drug response of individuals.
- *Association study*: study SNP patterns in diseased individuals and compare against SNP patterns in normal individuals.
- Many diseases associated with SNP profile.

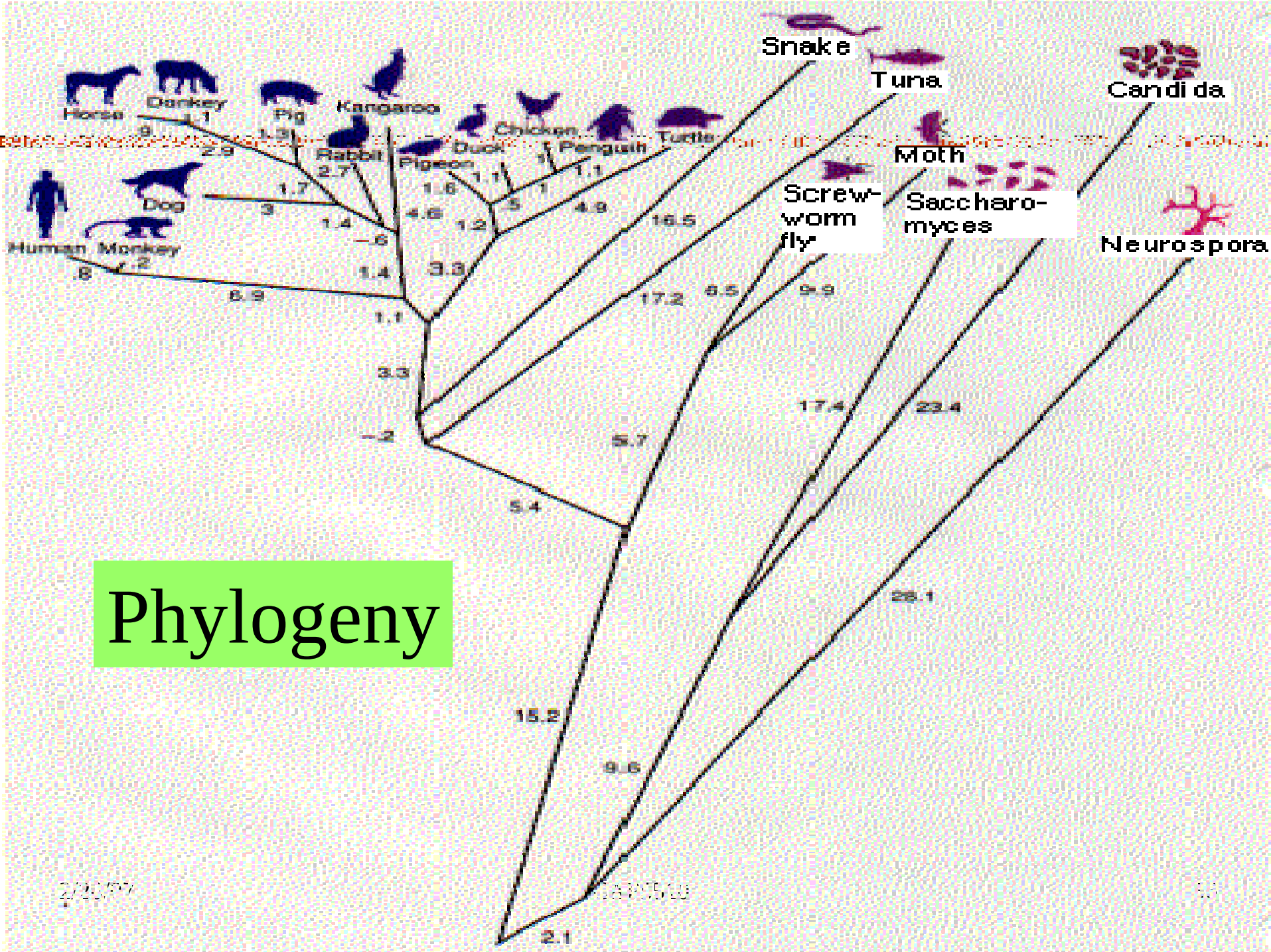
Theory of Evolution

- Charles Darwin
  - **1858-59:** *Origin of Species*
  - 5 year voyage of H.M.S. Beagle (1831-36)
  - Populations have variations.
  - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
  - Speciation arises by splitting of one population into subpopulations.
  - Gregor Mendel and his work (1856-63) on inheritance.


Millions of years
25
20
15
10
5
0
0.36
Pygmy chimp
0.44
0.37
Common chimp
0.36
0.84
Human
0.56
0.72
1.11
Gorilla
1.85
Orangutan
2.30
Gibbon

Dominant View of Evolution

- All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.
- Organization: **Directed Rooted Tree**; Existing species: Leaves; Common ancestor species (divergence event): Internal node; Length of an edge: Time.


Horse
Donkey
Pig
Kangaroo
Rabbit
Pigeon
Duck
Chicken
Penguin
Turtle
Dog
Human
Monkey
Snake
Tuna
Moth
Screw-worm fly
Saccharo-myces
Candida
Neurospora
.9
1.1
1.3
2.9
2.7
1.7
3
1.4
.6
1.4
.8
.2
6.9
1.1
3.3
2
1.6
4.6
1.2
1.1
1
.5
1
1.1
4.9
16.5
17.2
6.5
9.9
3.3
17.4
23.4
5.7
5.4
28.1
15.2
9.6
2.1

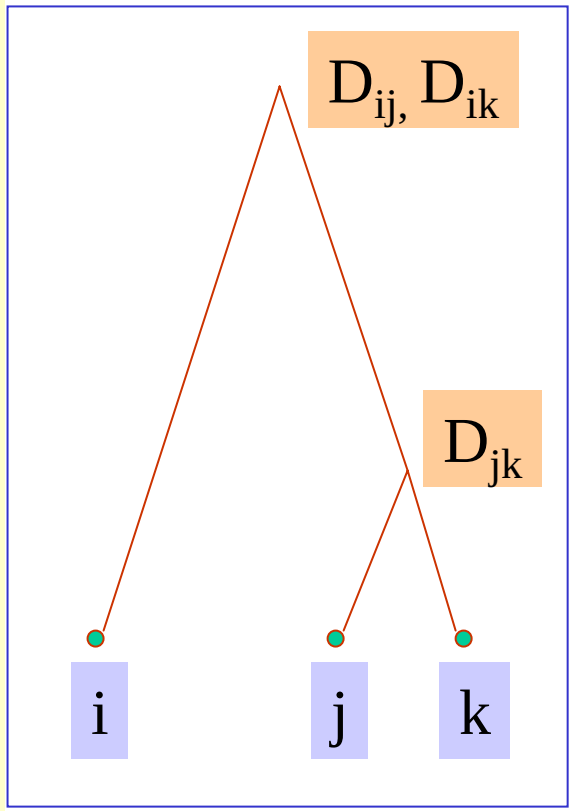
Phylogeny

Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Ultrametric

- An ultrametric tree:
  - decreasing internal node labels
  - distance between two nodes is label of least common ancestor.
- An ultrametric distance matrix:
  - Symmetric matrix such that for every i, j, k, there is tie for maximum of D(i,j), D(j,k), D(i,k)



Ultrametric: Assumptions

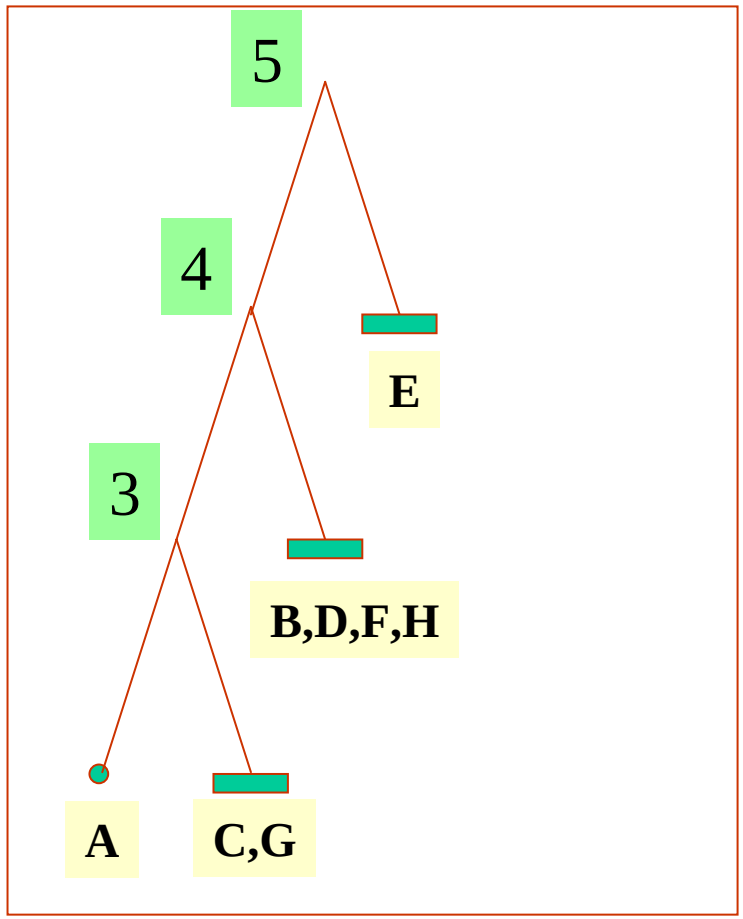
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

Ultrametric Data Sources

- Lab-based methods: hybridization
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: distance

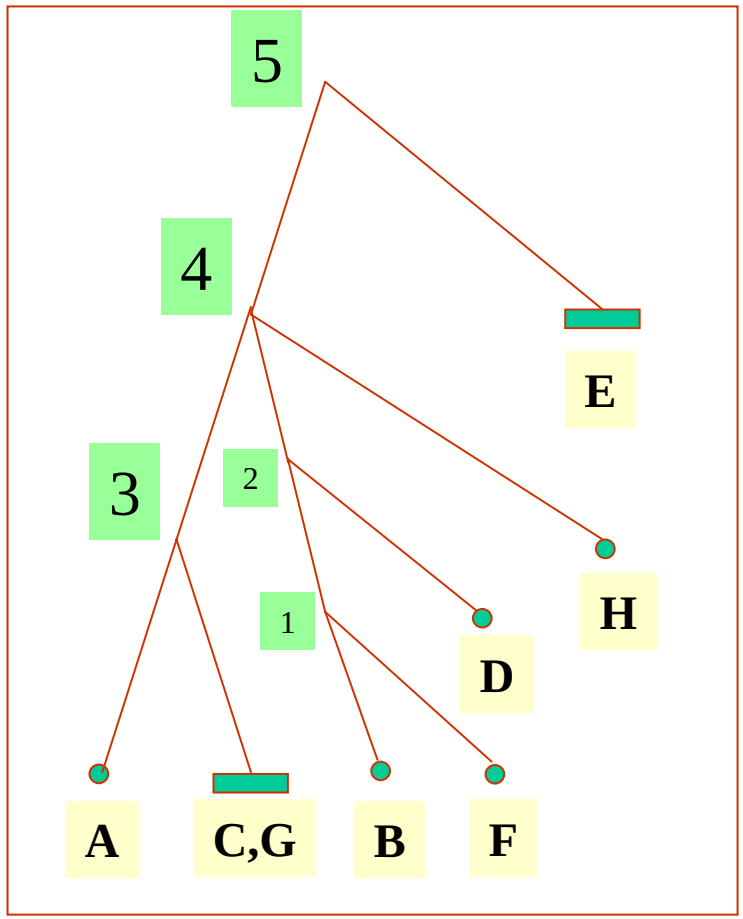
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B								
C								
D								
E								
F								
G								
H								



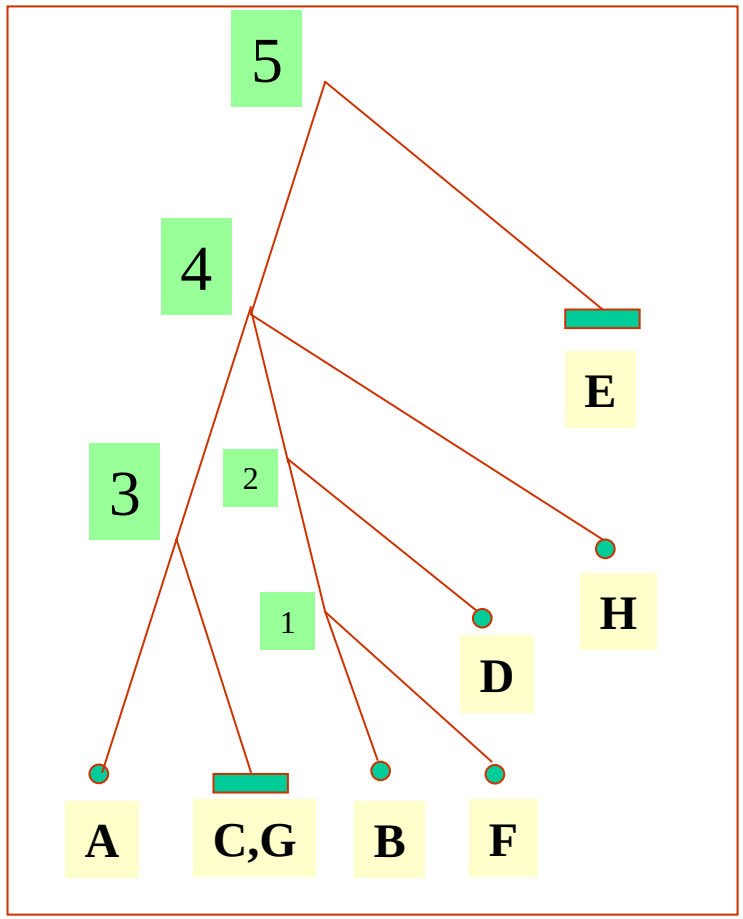
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C								
D								
E								
F								
G								
H								



Ultrametric: Distances Computed

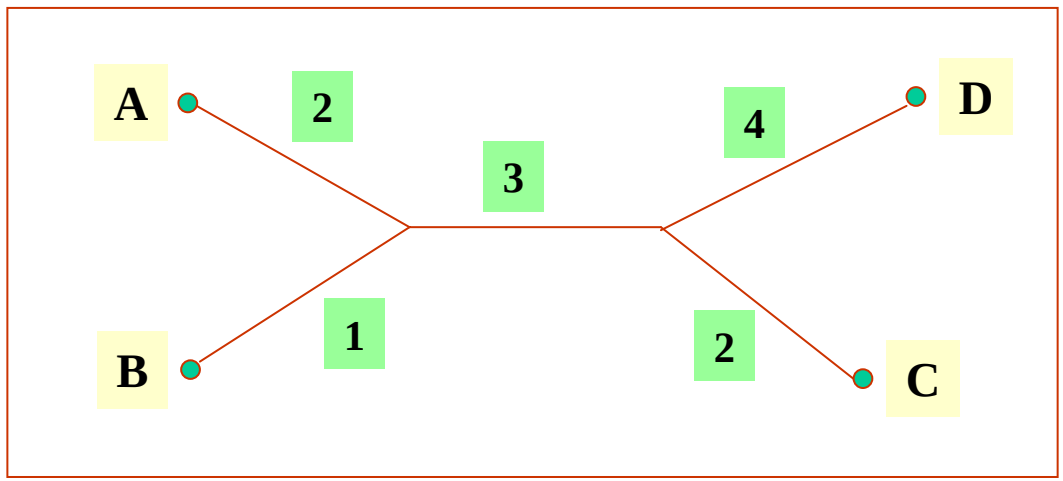
	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



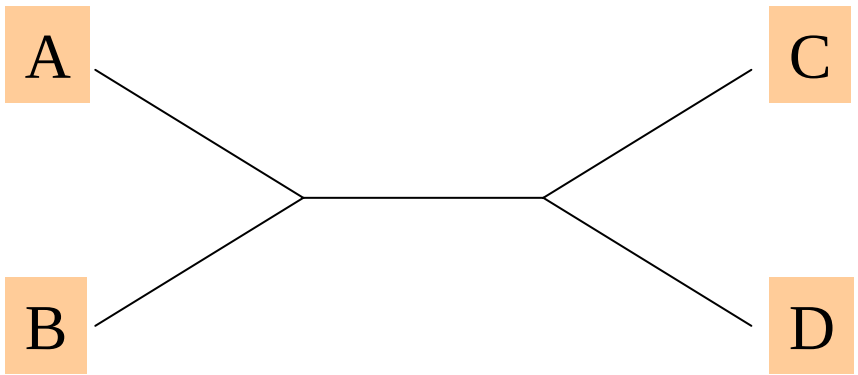
Additive-Distance Trees

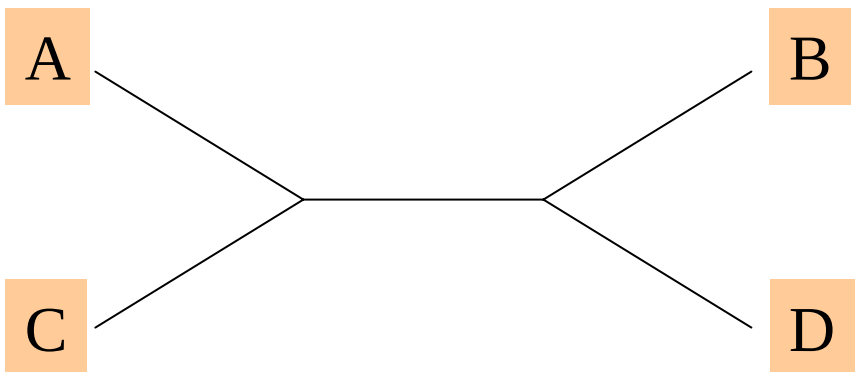
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

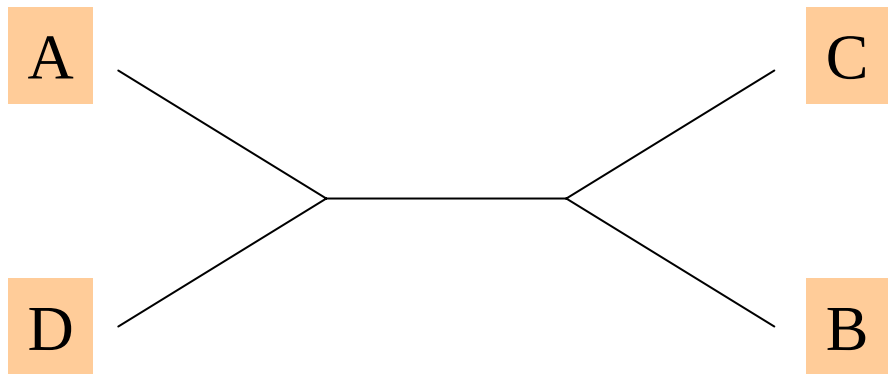
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

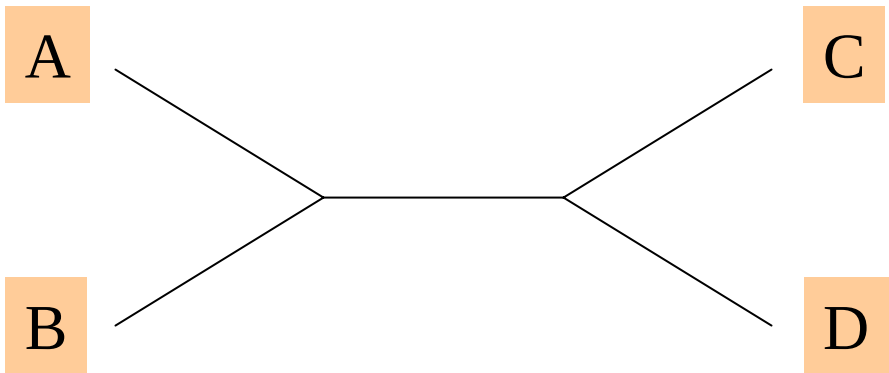






Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

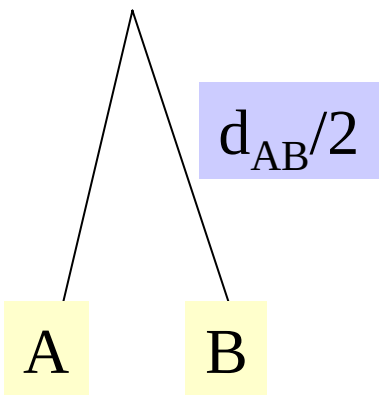


Unweighted pair-group method with arithmetic means (UPGMA)

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

	AB	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$$d_{(AB)C} = (d_{AC} + d_{BC}) /2$$



Transformed Distance Method

- UPGMA makes errors when rate constancy among lineages does not hold.
- Remedy: introduce an outgroup & make corrections

$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\sum_{k=1}^{n} D_{kO} \Big/ n \right)$$

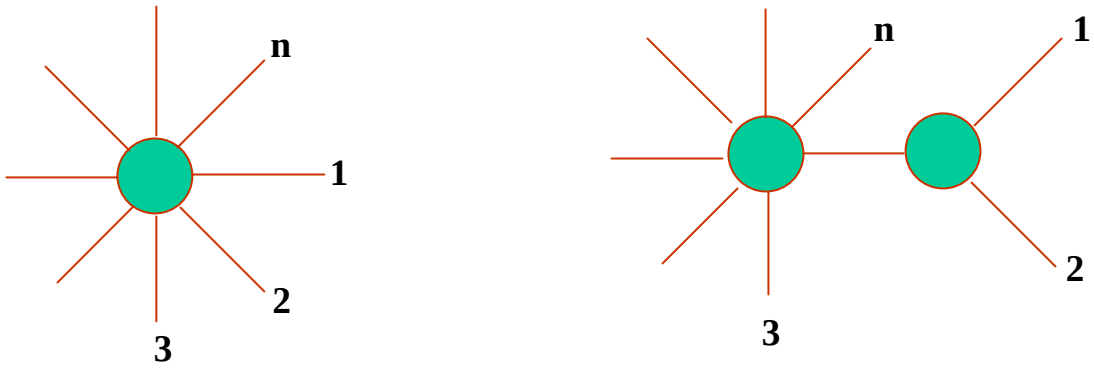
- Now apply UPGMA

Saitou & Nei: Neighbor-Joining Method

- Start with a star topology.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

Neighbor-Joining



$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

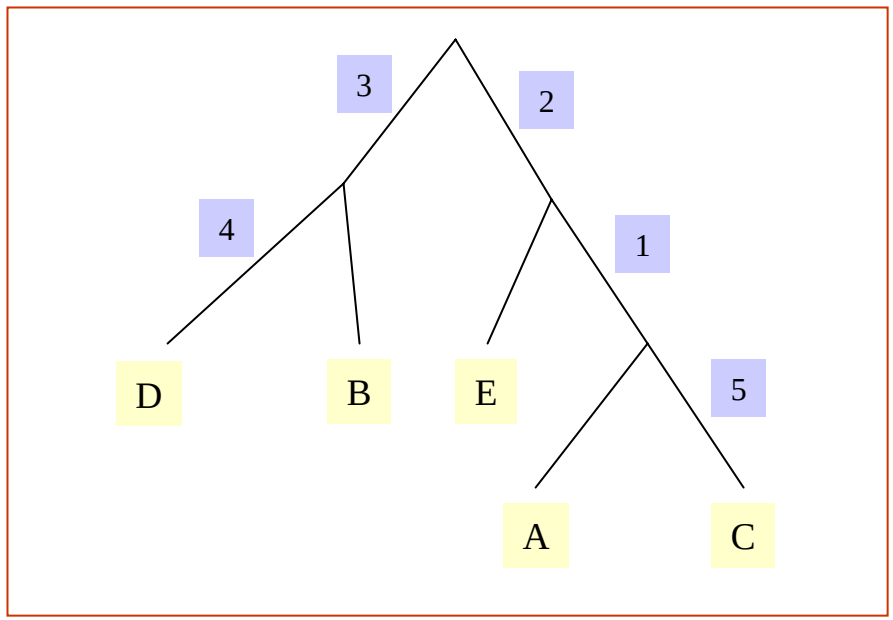
Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Character-based Methods

- Input: characters, morphological features, sequences, etc.
- Output: phylogenetic tree that provides the history of what features changed. [Perfect Phylogeny Problem]
- one leaf/object, 1 edge per character, path ⇔changed traits

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

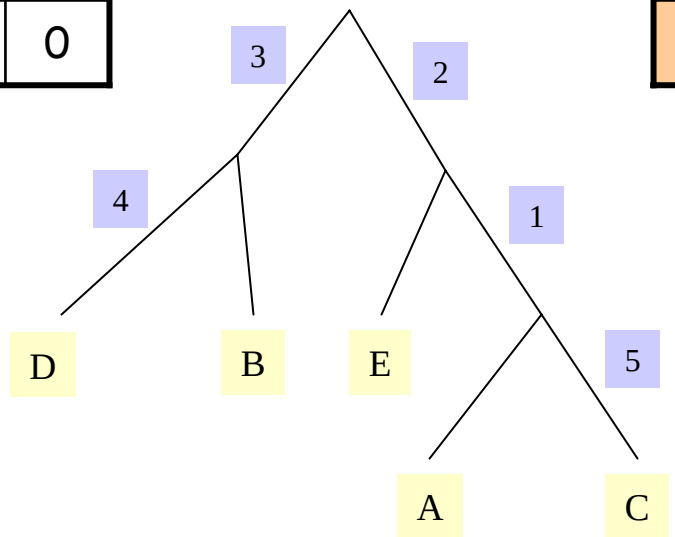


Example

- Perfect phylogeny does not always exist

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1



Maximum Parsimony

- Minimize the total number of mutations implied by the evolutionary history

Examples of Character Data

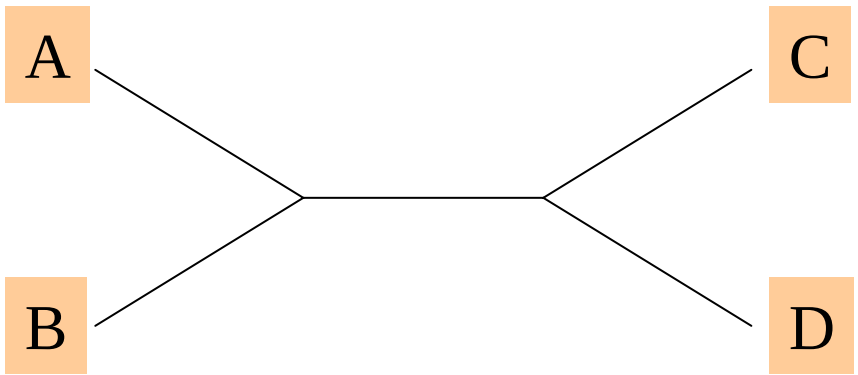
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

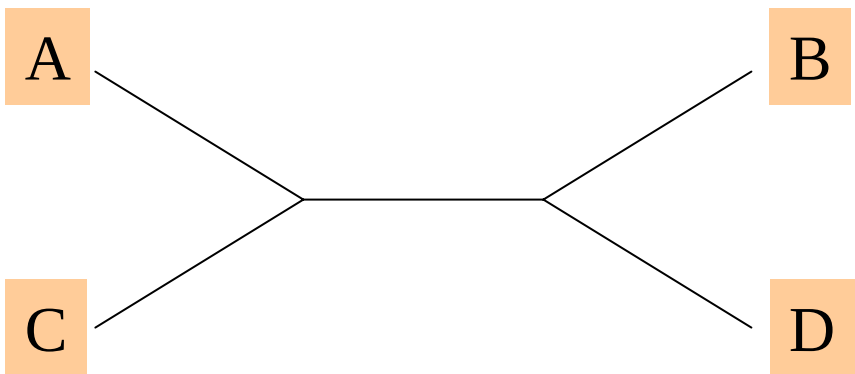
	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

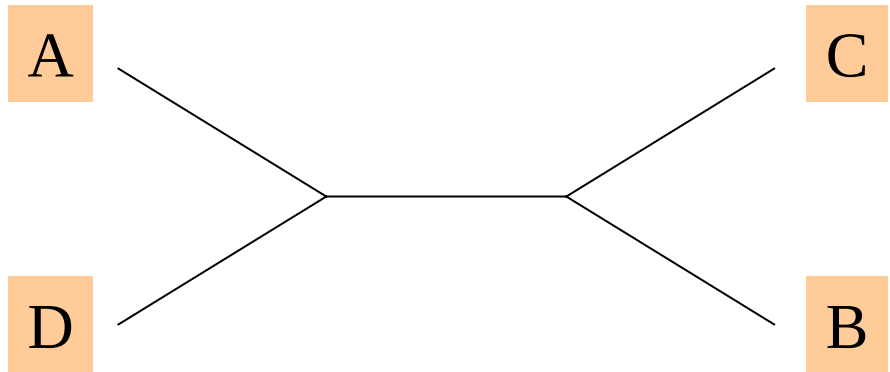
Maximum Parsimony Method: Example

	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Unrooted Trees on 4 Taxa







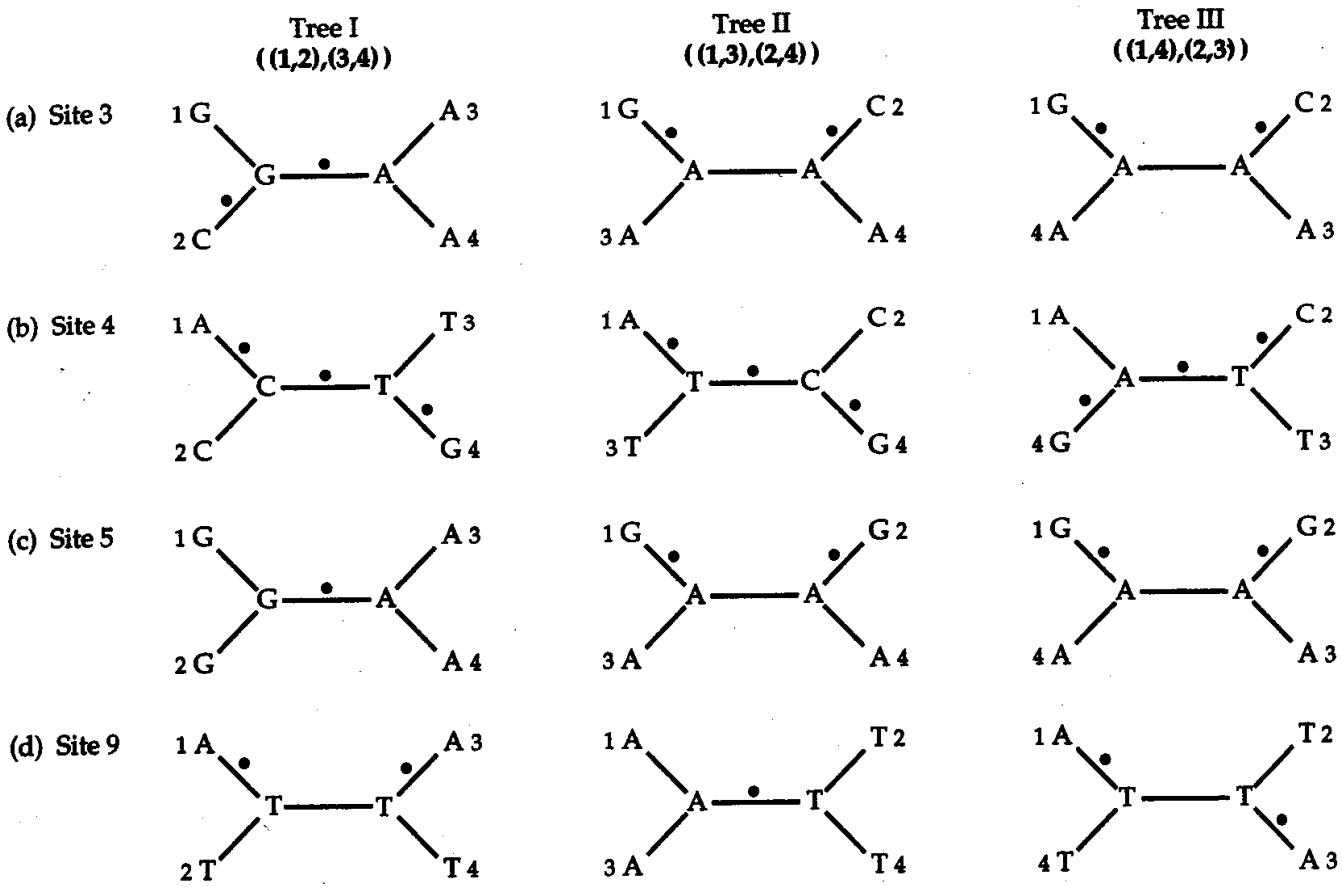


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III(d) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Inferring nucleotides on internal nodes

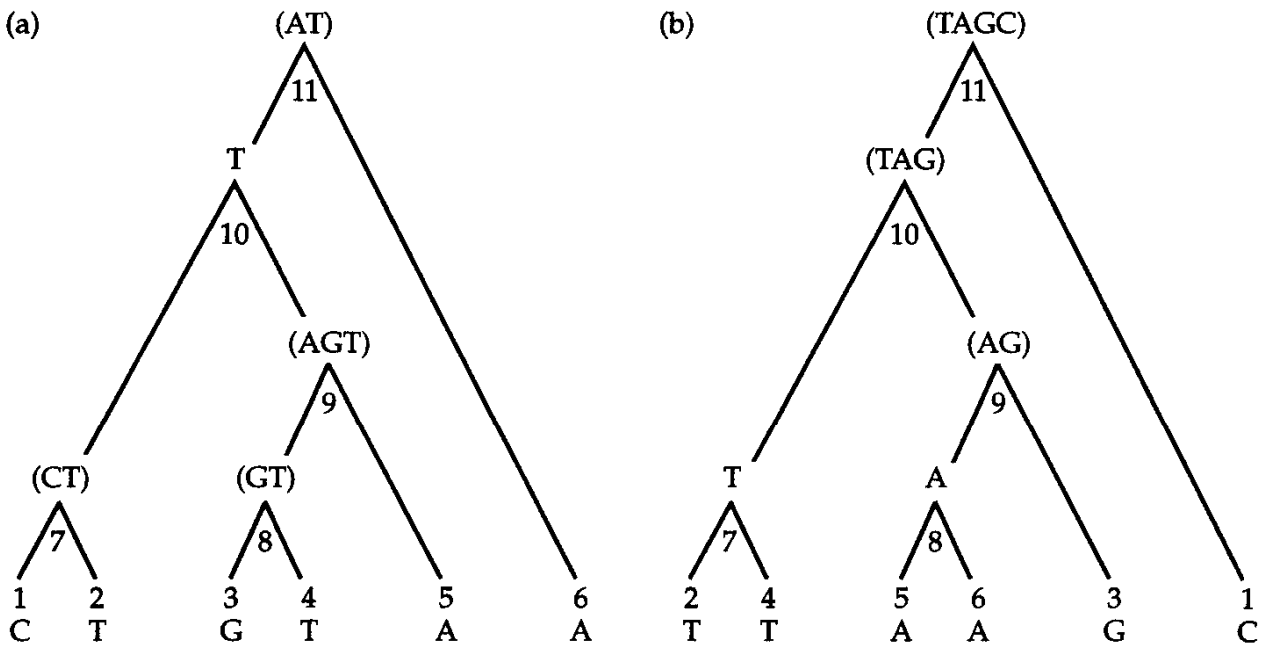


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).

Searching for the Maximum Parsimony Tree: Exhaustive Search

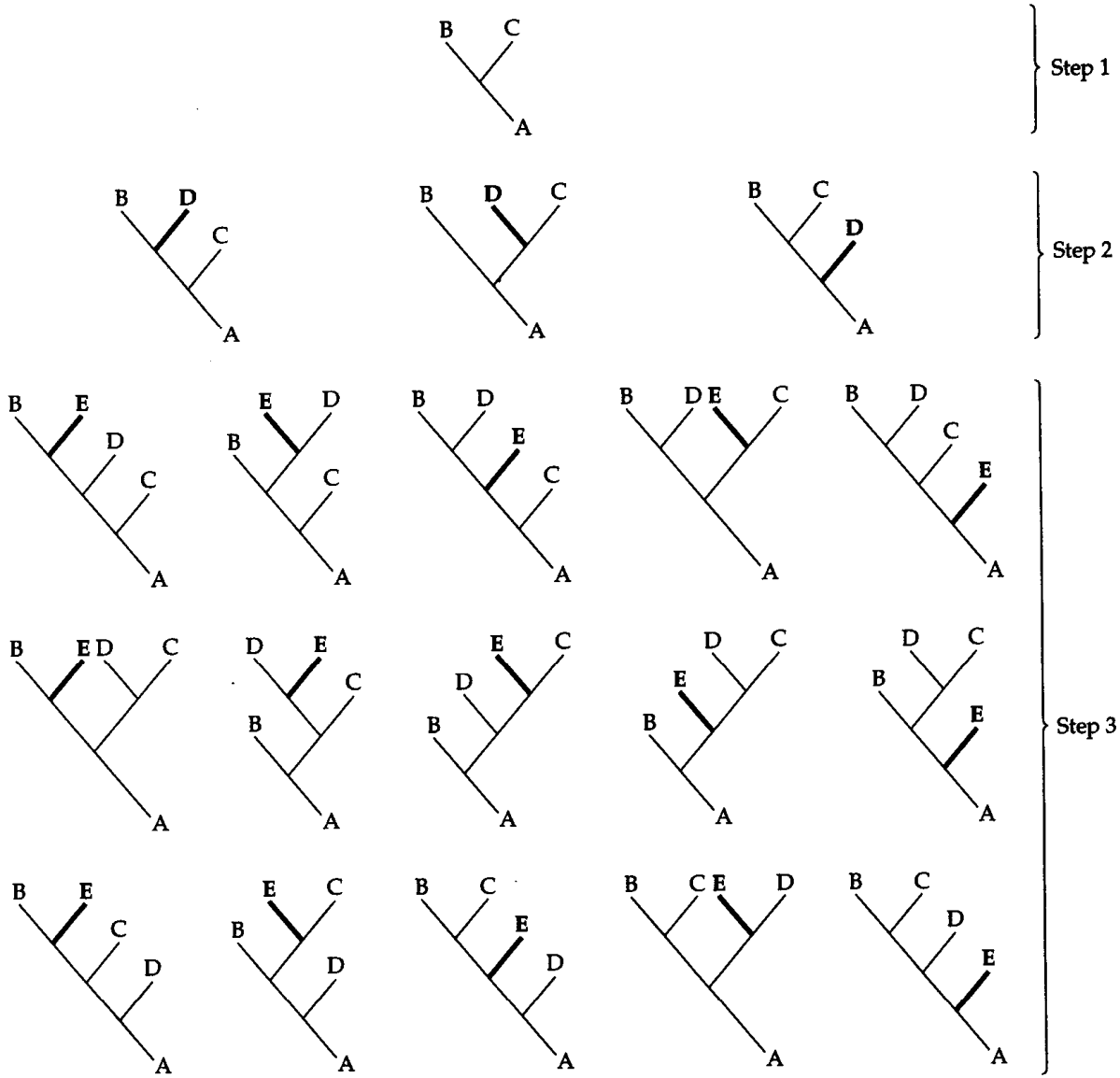
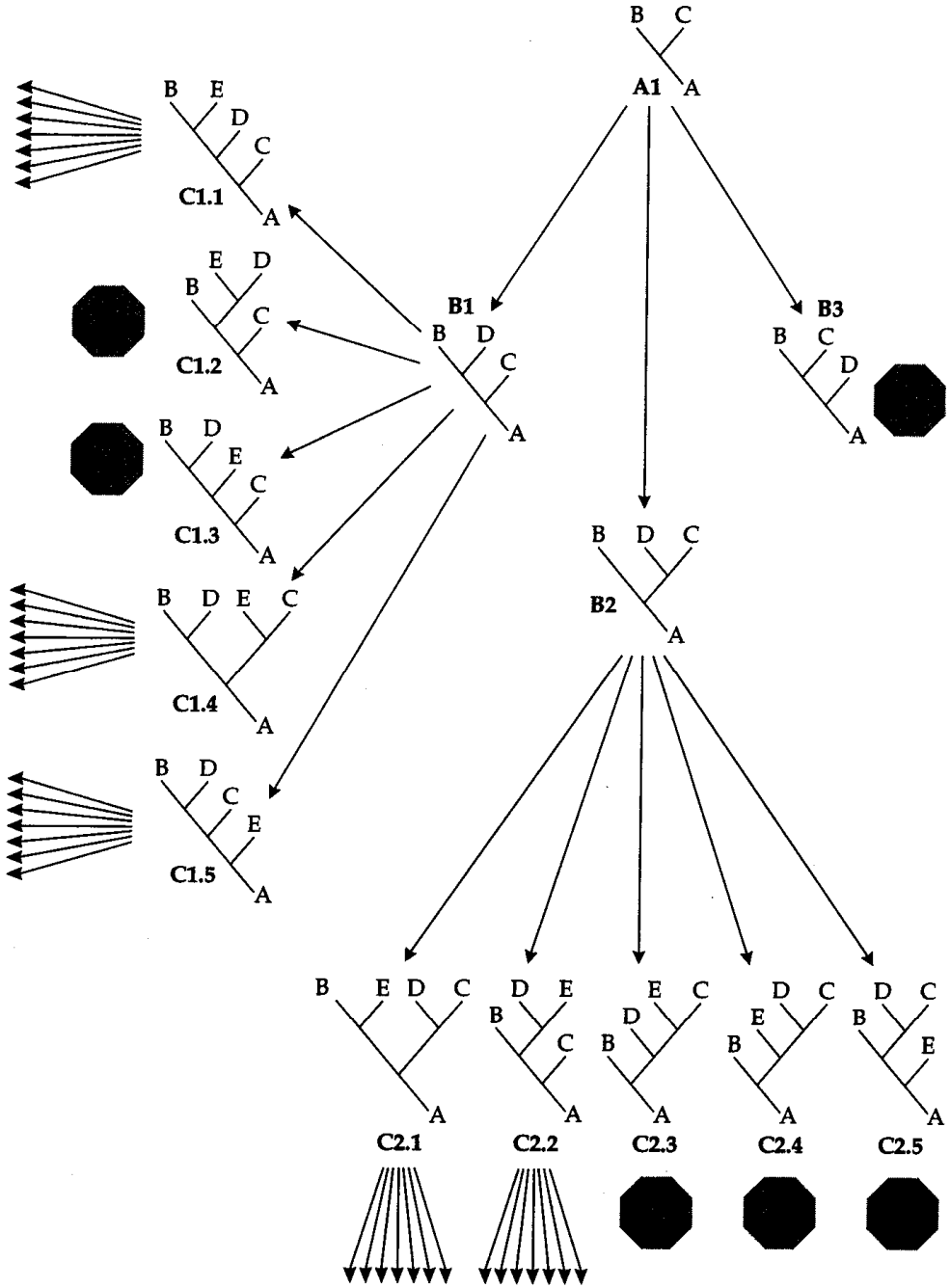


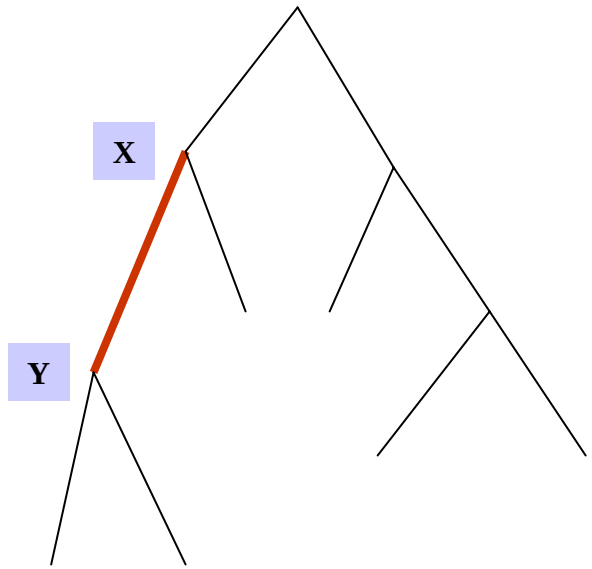
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four OTUs. In step 3, we add OTU E to each of the five branches of the three trees in step 2, thereby generating 15 unrooted trees. Additions of OTUs are shown as heavier lines. Modifed from Swofford et al. (1996).



Searching for the Maximum Parsimony Tree: Branch-&-Bound

Probabilistic Models of Evolution

- Assuming a model of substitution,
  - $Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$,
- Using this formula it is possible to compute the likelihood that data D is generated by a given phylogenetic tree T under a model of substitution. Now find the tree with the maximum likelihood.



- Time elapsed? Δ
- Prob of change along edge? $Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$
- Prob of data? Product of prob for all edges

(a)

	1	2	3	4	5	6	7	8	9	...*n*
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N

(b)

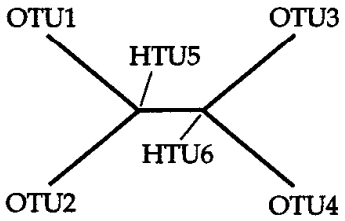
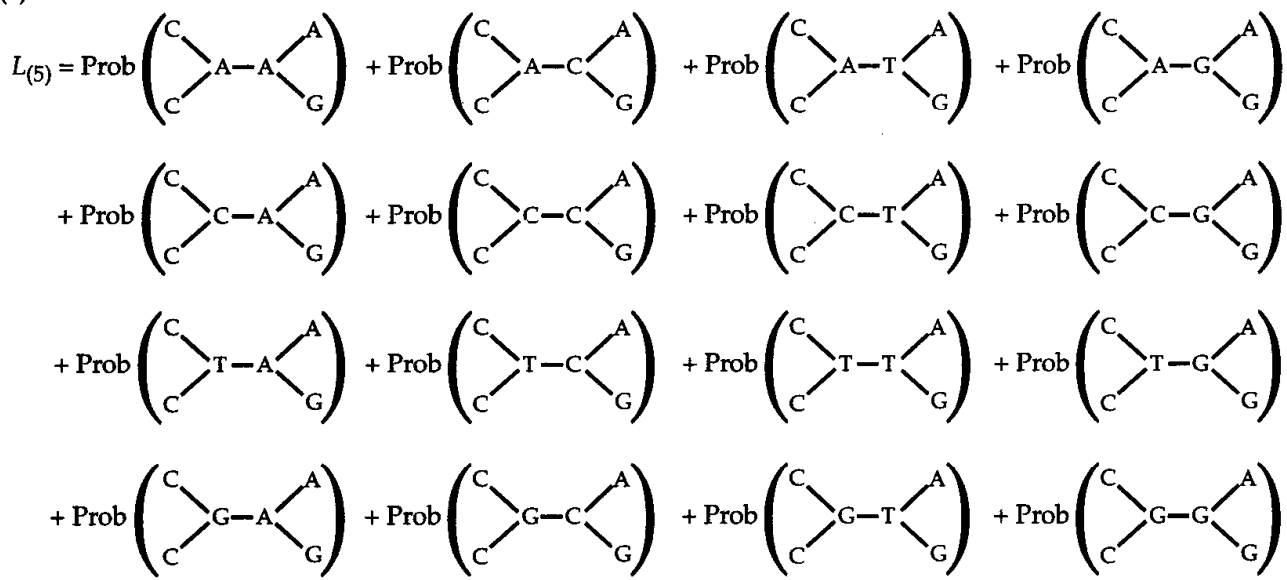


FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length *n*. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all *n* sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).

(c)



(d) $$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \ldots \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

(e) $$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \ldots + L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)}$$

Computing Maximum Likelihood Tree

Genomics

- Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker: uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)

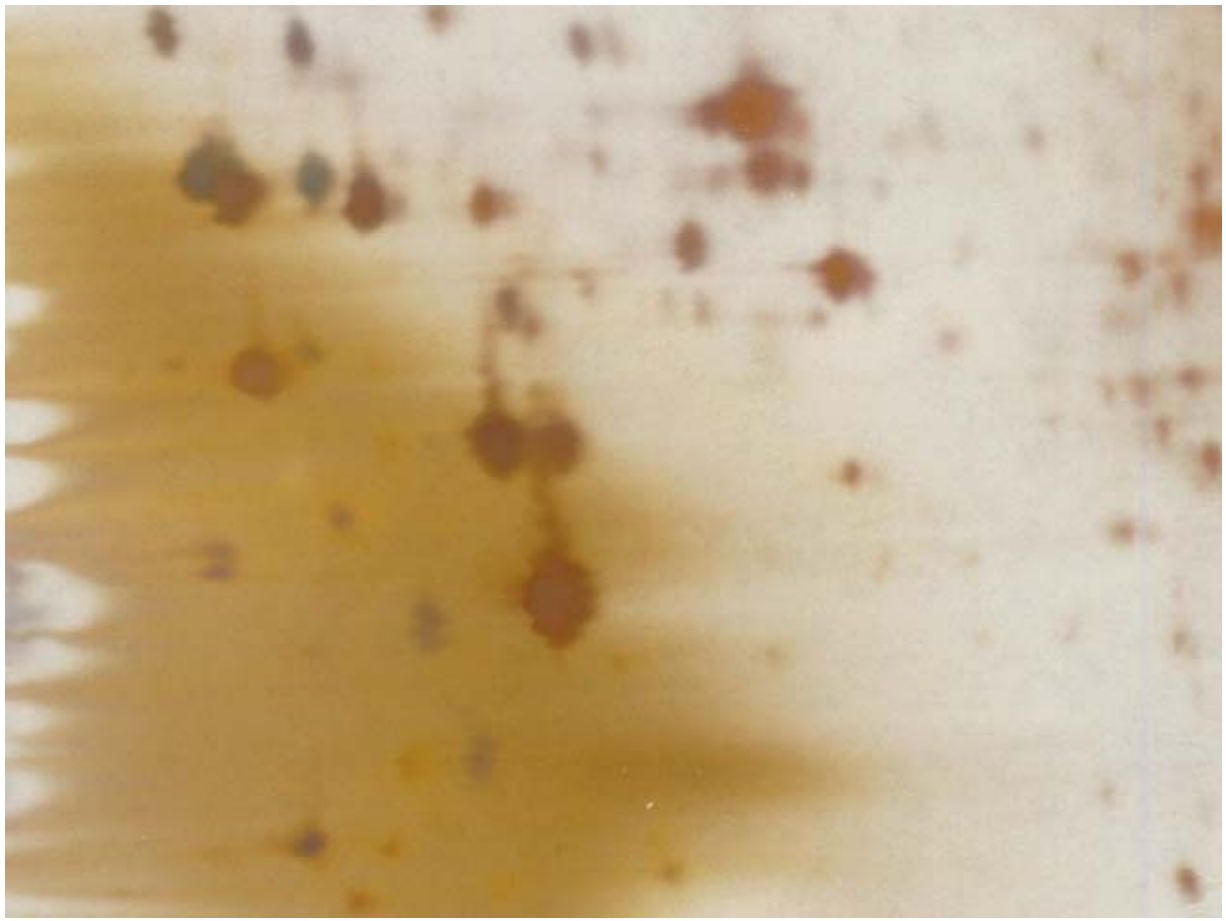
Genomics (Cont'd)

- Gene Expression
  - Microarray experiments & analysis
    - Probe design (CODEHOP)
    - Array image analysis (CrazyQuant)
    - Identifying genes with significant changes (SAM)
    - Clustering

Proteomics

- Study of all proteins in a genome, or comparison of whole genomes.
  - Whole genome annotation & Functional proteomics
  - Whole genome comparison
  - Protein Expression: 2D Gel Electrophoresis

2D Gel Electrophoresis



Other Proteomics Tools

From ExPASy/SWISS-PROT:

- **AACompIdent** identify proteins from aa composition

[Input: aa composition, isoelectric point, mol wt., etc. Output: proteins from DB]

- **AACompSim** compares proteins aa composition with other proteins
- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins
- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins
- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.
- **PeptideMass** theoretical mass fingerprint for a given protein.
- **GlycoMod** predicts oligosaccharide modifications from mass difference
- **TGREASE** calculates hydrophobicity of protein along its length

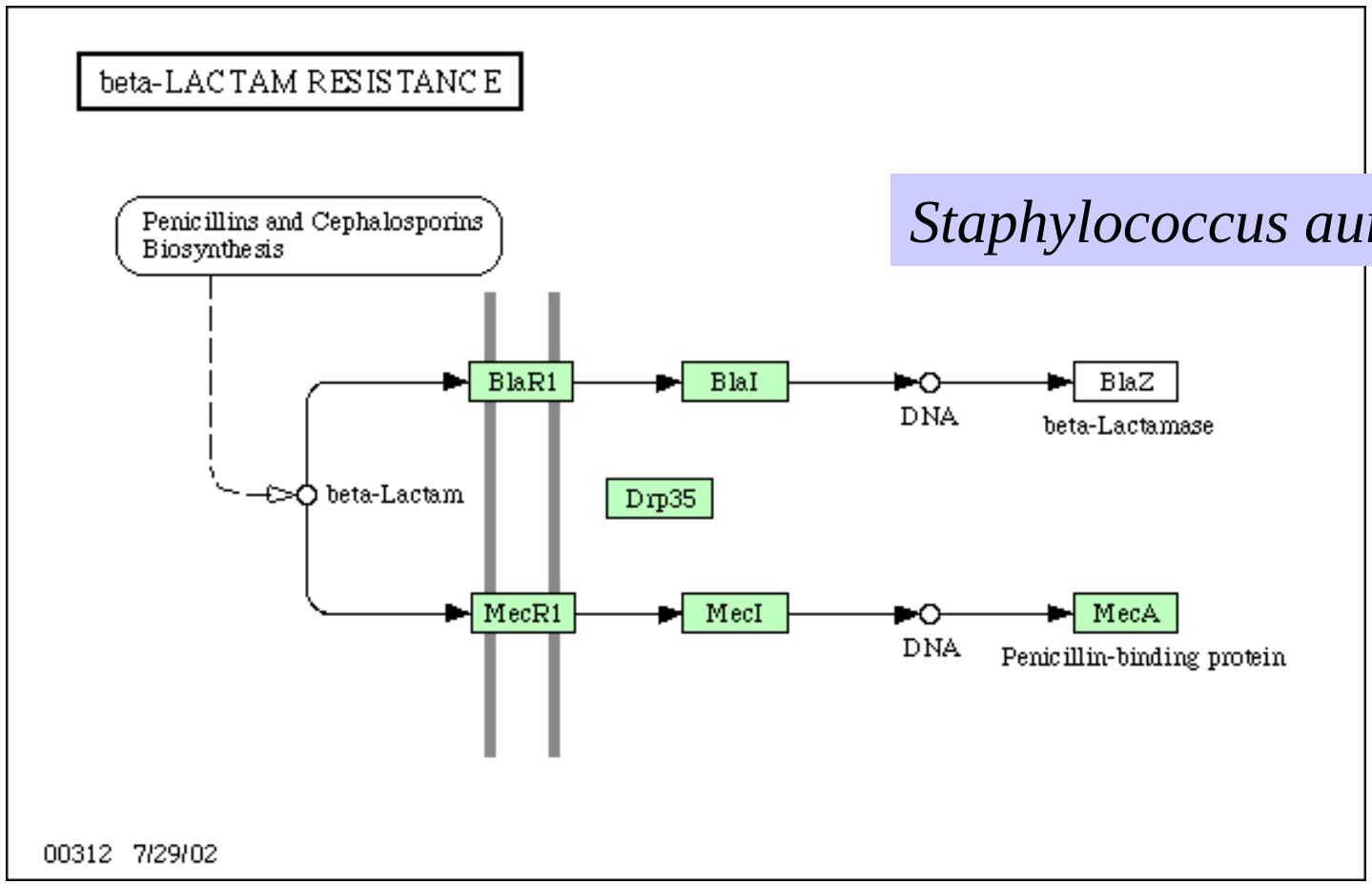
Databases for Comparative Genomics

- PEDANT useful resource for standard questions in comparative genomics. For e.g., *how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.*
- COGs Clusters of orthologous groups of proteins.
- MBGD Microbial genome database searches for homologs in all microbial genomes

Gene Networks & Pathways

- Genes & Proteins act in concert and therefore form a complex network of dependencies.

Pathway Example from KEGG



Staphylococcus aureus

Pseudomonas aeruginosa